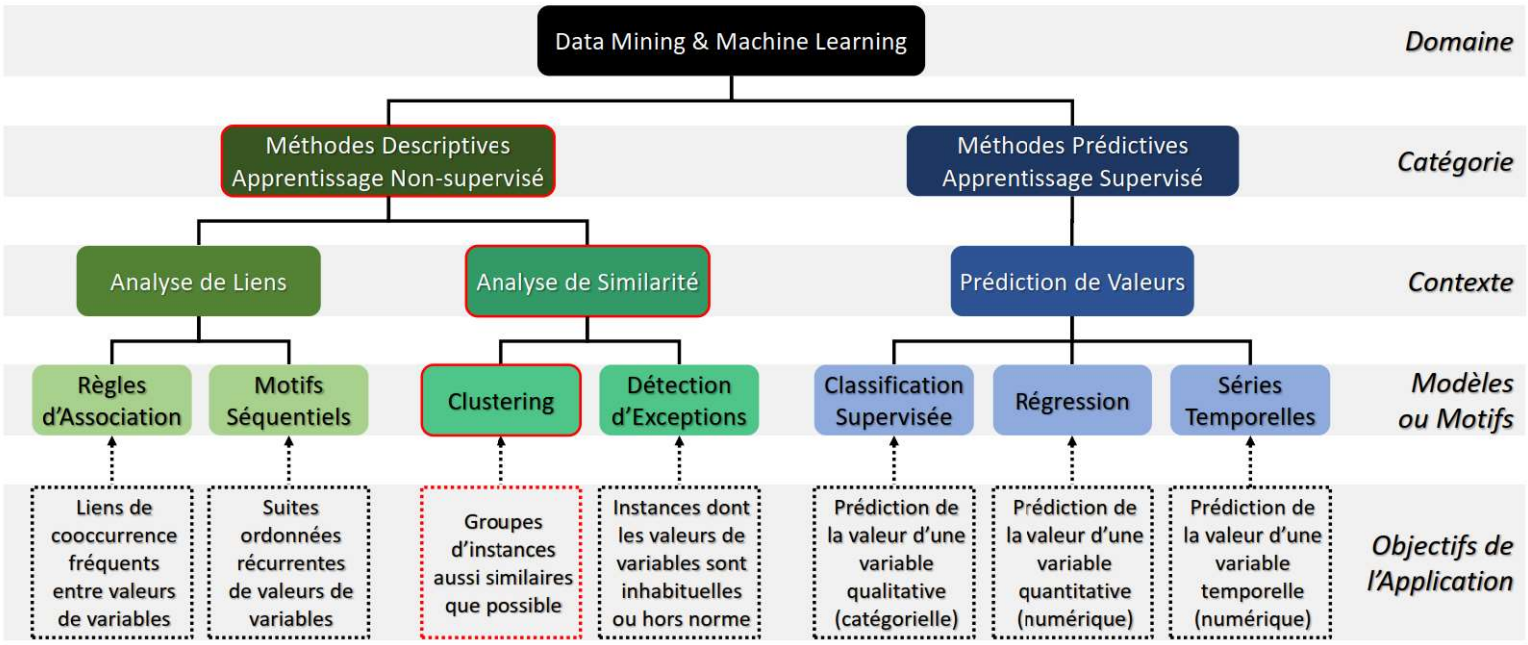
Introduction au Clustering de Données

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

Taxonomie des Méthodes d'Extraction de Modèles de Connaissances



Objectifs du Cours

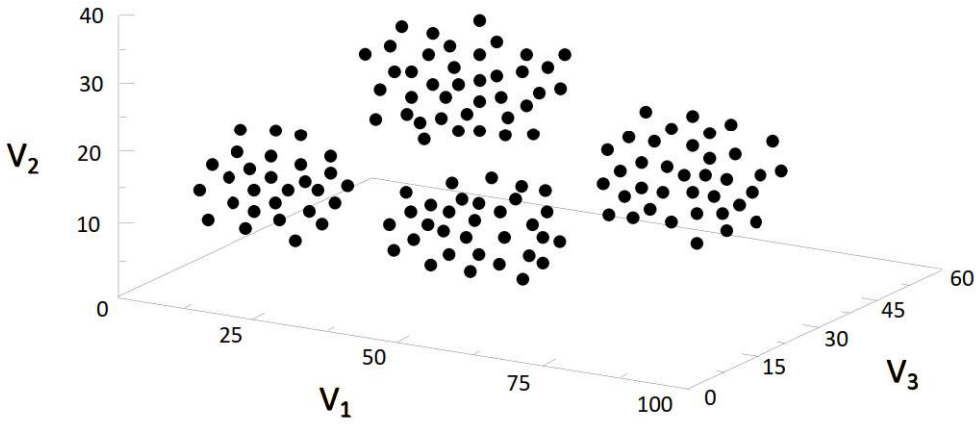
- Comprendre la notion d'apprentissage non-supervisé
- La relier à la notion de découverte de structures dans l'espace des données
- Connaître les approches algorithmiques de clustering
  - Par partitionnement, hiérarchique, basée sur la densité, basée sur la décomposition de l'espace des données en grilles, par modèles de clusters ou concepts, par consensus (*Ensemble clustering*)
- Comprendre la notion de similarité, liée à la notion mathématique de distance, qui est subjective mais centrale dans cette problématique
- Être en mesure de :
  1. Construire un espace des données multi-dimensionnel
  2. Définir une mesure de similarité dans cet espace des données
  3. Choisir l'algorithme à utiliser en fonction des données en entrée
  4. Choisir des paramétrages adéquats pour les algorithmes testés

Qu'est ce que le Clustering ?

- Un cluster est un groupe d'instances (lignes de la matrice de données) qui sont autant que possible :
  - Similaires entre elles au sein d'un même groupe
  - Différentes d'un groupe à l'autre
- Le clustering est le processus de classification des instances dans différents groupes
- Contexte non-supervisé : aucune variable cible ou de classes
  - Nous n'avons donc aucune connaissance préalable du nombre et du type de clusters « naturels » dans l'espace de données
  - Processus subjectif qui vise à découvrir des structures dans l'espace de données afin de révéler des groupes de données cohérents
- Terminologie : segmentation, apprentissage non supervisé, partitionnement des données

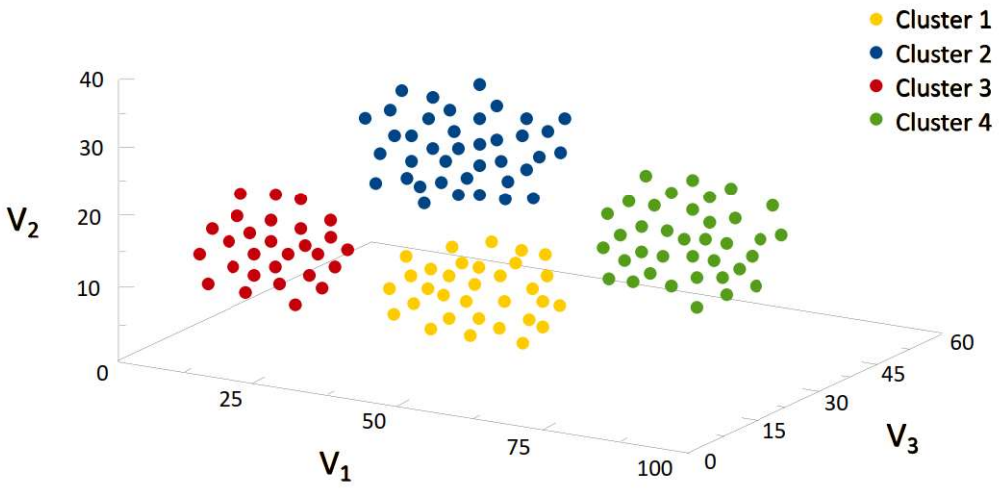
Exemple : Espace Multi-dimensionnel

- Espace tri-dimensionnel des données
  - Dimensions : variables V_1, V_2, V_3
  - Chaque instance de l'ensemble de données est représentée par un point



Exemple : Clustering

- Quatre clusters «naturels» : quatre régions denses dans l'espace des données séparées par des régions faiblement peuplées



Qu'est ce qu'un Bon Clustering?

- Évaluer la qualité des groupes découverts
  - Minimiser la variabilité intra-clusters (i.e. haute similarité des instances dans les clusters)
  - Maximiser la variabilité inter-clusters (i.e. faible similarité entre instances de différents clusters)
- Similarité entre deux instances
  - Évaluée en comparant les valeurs des variables pour ces instances
  - Estimée par le calcul d'une distance entre elles
- La qualité du résultat du clustering dépendra de :
  - La mesure de distance utilisée
  - La configuration algorithmique choisie pour la mettre en œuvre

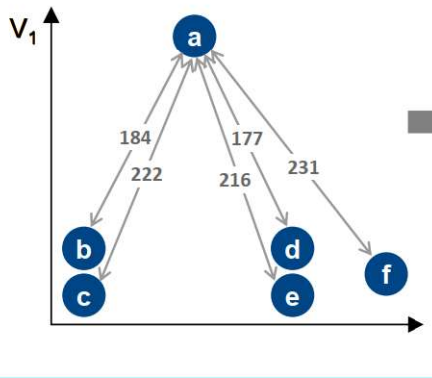
Mesure de Distance : Définition

- Soit X et Y deux vecteurs (instances)
- Une fonction d() est une mesure de distance si et seulement si d(X, Y) vérifie les propriétés suivantes (Anderberg, 1973)
  i. Non négativité : $d(X, Y) \geq 0$
  ii. Réflexivité : $d(X, X) = d(Y, Y) = 0$
  iii. Commutativité : $d(X, Y) = d(Y, X)$
  iv. Inégalité triangulaire : $d(X, Y) \leq d(X, W) + d(W, Y)$
- La définition des mesures de distance dépend du type des variables dans les données (numériques, binaires, etc.)
- Il est difficile de définir la notion de « suffisamment similaire » pour inclure deux instances au sein du même groupe : il y a généralement une part de subjectivité dans la décision

Matrice de Distance : Définition

- Le calcul de la distance pour chaque paire d'instances de l'ensemble de données génère une matrice de distances
- Exemple : matrice de distance Euclidienne pour un ensemble de six instances (a, b, c, d, e, f) et deux dimensions numériques (V_1 et V_2)

Espace des données Bi-dimensionnel



Calcul des distances inter-instances

Matrice de distance Euclidienne

	a	b	c	d	e	f
a	0					
b	184	0				
c	222	45	0			
d	177	123	129	0		
e	216	128	121	46	0	
f	231	203	201	83	81	0

Mesure de Distance : Types de Variables

- La mesure de distance est définie en fonction des types de variables (sémantique de chaque variable et non son encodage)
- Numérique : valeurs continues
  - Ex : Température $\in \mathbb{Z}$, âge $\in \mathbb{N}$, vitesse $\in \mathbb{R}$
  - Les échelles non-linéaires (logarithmiques $\beta.\log(\alpha.n)$ ou exponentielles $\beta.e^{(\alpha.v)}$) sont des cas à part (traitées comme des valeurs ordinales)
- Binaire : deux valeurs possibles
  - Ex : Genre $\in$ {H, F}, Marié $\in$ {vrai, faux}, Actif $\in$ {0, 1}
- Catégorielle (nominale) : liste de valeurs discrètes possibles
  - Ex : Couleur $\in$ {bleu, vert, …}, Numéro Dépt. $\in$ [01, 95]
- Ordinale : liste de valeurs discrètes ordonnées possibles
  - Ex : Niveaux $\in$ {faible, moyen, élevé}, Classement $\in$ {1er, 2e, 3e, …}

Mesure de Distance : Variables Hétérogènes

- Soient $X = \{X_1, ..., X_p\}$ et $Y = \{Y_1, ..., Y_p\}$ deux instances définies par p variables hétérogènes (binaires, numériques, catégorielles, etc.)
- Formule pondérée de combinaison des différentes mesures de distances, par exemple :

$$d(X, Y) = \frac{\sum_{i=1}^{i=p} \delta_i(X, Y)\, d_i(X, Y)}{\sum_{i=1}^{i=p} \delta_i\,(X, Y)}$$

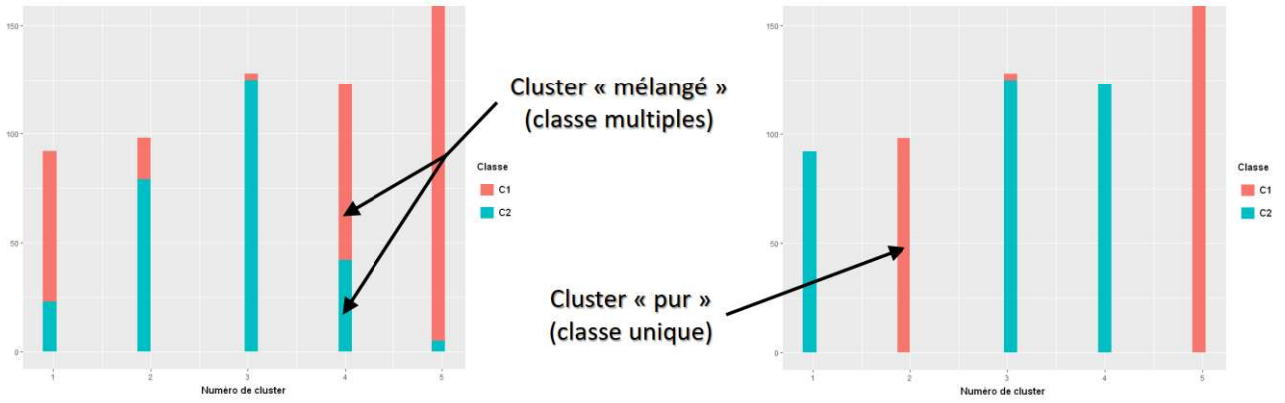
- $d_i(X,Y)$ est la mesure de distance entre X et Y pour la variable V_i
- $\delta_i(X,Y)$ est une valeur binaire 1 ou 0
  - Pondère (ignoré ou non) le calcul de $d_i(X,Y)$ pour tenir compte :
    - Des variables asymétriques : une combinaison spécifique de valeurs est ignorée, e.g. si $X_i=0$ et $Y_i=0$
    - Des valeurs manquantes : X_i=NA ou/et Y_i=NA

Mesure de Distance : Variables Hétérogènes

- La valeur de $d_i(X,Y)$ est calculée selon le type de la variable V_i
  - Variable binaire ou discrète : si $X_i = Y_i$ alors $d_i(X,Y) = 0$ sinon $d_i(X,Y) = 1$
  - Variable numérique continue : utiliser une mesure de distance normalisée (Euclidienne, Manhattan, Mahalanobis, etc.)
  - Variable ordinale ou numérique sur une échelle non linéaire (exp/log) :
    1. Ordonner les valeurs par ordre croissant
    2. Calculer les rangs des valeurs
    3. Normaliser les rangs en calculant leur z-score $\in [0.0, 1.0]$
    4. Traiter les valeurs résultantes comme numériques continues
- L'argument $\delta_i(X,Y)$ prend la valeur :
  - 0 si une des valeurs X_i ou Y_i est manquante
  - 0 si $X_i = Y_i = 0$ et V_i est binaire asymétrique (combinaison de valeurs 0 ignorée, par ex. symptôme médical absent chez les deux patients)
  - 1 sinon

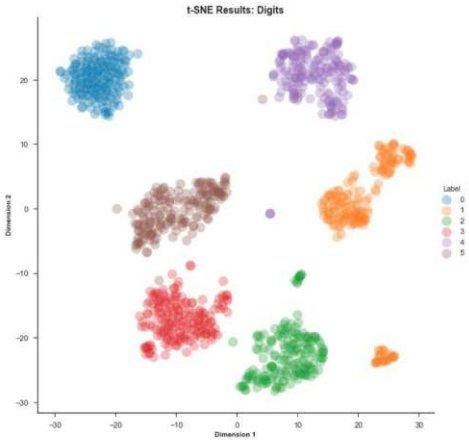
Évaluation des Clusters : Évaluation Externe

- **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

- Bibliographie
  - C. C. Aggarwal & C. K. Reddy. *Data Clustering: Algorithms and Applications*. CRC Press, August 2013.
  - G. Gan, C. Ma & J. Wu. *Data Clustering: Theory, Algorithms, and Applications*. SIAM Publisher, July 2007.
  - M. J. Zaki & W. Meira. *Data Mining and Analysis – Fundamental Concepts and Algorithms*. Cambridge University Press, 2014.
- Web sites
  - KD Nuggets: Business Analytics, Big Data, Data Mining, Data Science, and Machine Learning. http://www.kdnuggets.com/
  - R and Data Mining: Book, documents, examples, tutorials and resources on R and data mining. http://www.rdatamining.com/
  - CRAN Task View: Cluster Analysis & Finite Mixture Models. https://cran.r-project.org/web/views/Cluster.html