Programmation de Modèles Linguistiques 2, L6SOPRG L3

Caroline Koudoro-Parfait
caroline.parfait@sorbonne-universite.fr

Observatoire des Textes des Idées et des Corpus - Obtic,
Sorbonne Center for Artificial Intelligence - SCAI,
Sens Textes Informatiques Histoire - STIH EA 4509, Sorbonne Université

Plan du cours

Cluster ou Partitionnement de données

Le Clustering une discipline de Machine learning

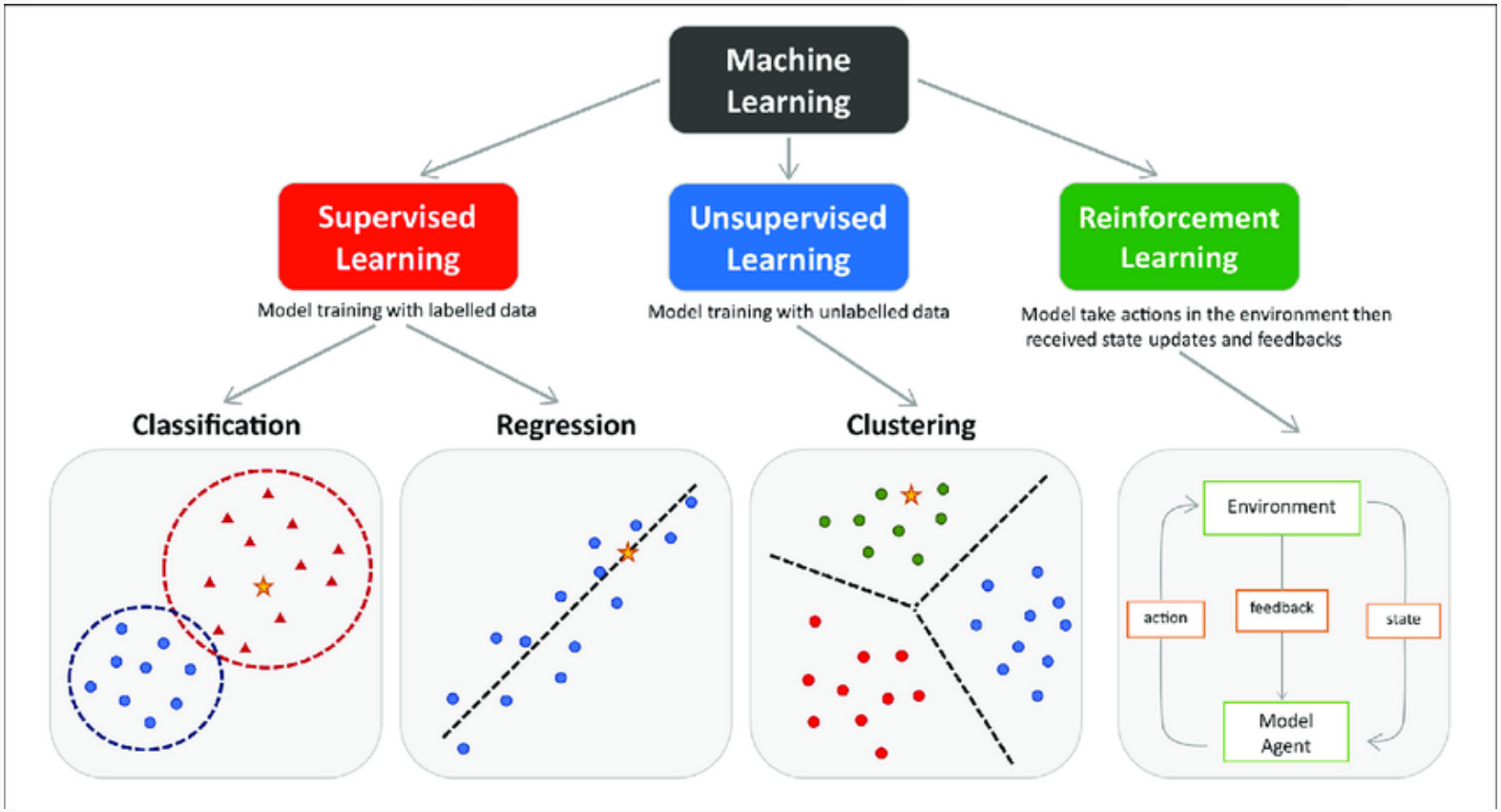

Figure 1 – https://www.frontiersin.org/journals/pharmacology/articles/10.3389/fphar.2021.720694/full, [?]

Définition [1]

Général *Groupement d'un petit nombre d'objets.*

Linguistique *groupe consonantique, il correspond à une succession d'au moins deux consonnes dans un mot. Par exemple, dans fraise ou tigre. On retrouve parfois des clusters plus complexes.*

Informatique 1 *grappe de serveurs sur un réseau*

Informatique 2 *une base de données distribuée dans des grappes de serveurs - MySQL Cluster*

Informatique 3 *≪ Data clustering ≫ désigne l'analyse de* **partitionnement de données**.

---

1. `https://fr.wikipedia.org/wiki/Cluster`,Wikipédia

Le partitionnement de données [2]

* Méthode de classification non supervisée,
* algorithmes d'apprentissage
* regrouper des données non étiquetées selon des propriétés similaires
* Isoler des schémas/familles

2. https://dataanalyticspost.com/Lexique/clustering/

à quoi ça sert ?

* En machine learning

  ➜ préparer l'application d'algorithmes d'apprentissage supervisé → **KNN**.

  ➜ utilisé lorsqu'il est coûteux d'étiqueter le données.

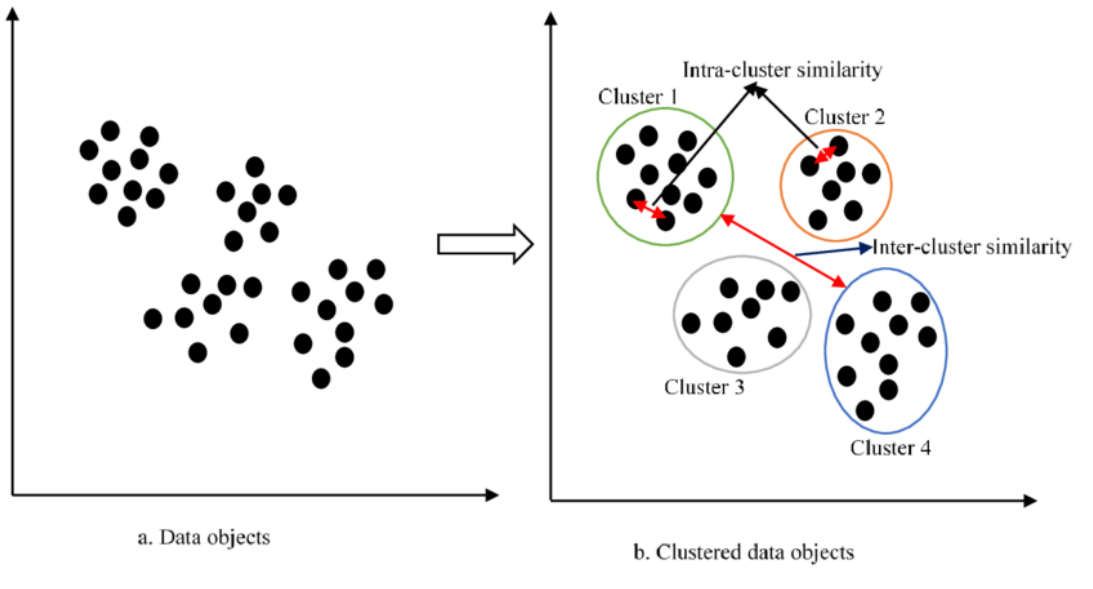

Figure 2 – `https://www.researchgate.net/figure/Clustering-example-with-intra-and-inter-clustering-illustrations_fig1_344590665`, [**?**]

à quoi ça sert ?

* En TAL

  ➜ Regrouper des textes

  ➜ Regrouper des mots

➥ selon :

  ➜ des caractéristiques linguistique commune (par. ex : quantité de verbes, noms, verbes etc.)

  ➜ leur sens → partitionnement sémantique

Logique du plus proche voisin

Algorithme du plus proche voisin, *K-nearest neighbors* (KNN ou K-NN)[3], [4]

➜ méthode d'apprentissage supervisé,

➜ utilisé pour la classification et la régression,

➜ pas de phase d'apprentissage ➞ Lazy Learning. L'algo. généralise directement à partir du jeu de données.

➜ Observe la similarité des K voisins les plus proches. K = nombre entier

➥ *"dis moi qui sont tes voisins, je te dirais qui tu es. . . "*

3. https://fr.wikipedia.org/wiki/M%C3%A9thode_des_k_plus_proches_voisins
4. https://mrmint.fr/introduction-k-nearest-neighbors

Les limites du partitionnement de données

* Pour les mêmes données, peuvent être utilisées :

  ➜ différentes métriques,

  ➜ différentes représentations des données

➜ il peut y avoir des variations/différents regroupements dans les clusters en sortie

* Choisir la méthode dont vous partitionnez les données en considérant :

  * les résultats attendus,

  * l'utilisation prévue des données

Le clustering en TAL : quelles méthodes

* Méthodes et Algorithmes :

  ➜ hiérarchique : dendogrames

  ➜ centroïde : K-mean, Affinity Propagation

  ➜ densité : « *density-based spatial clustering of applications with noise* » - DBSCAN

  ➜ maximisation de l'espérance (EM) : outils mathématiques probabilistes (Loi Gausse - *Gaussian Mixture model*[5])

5. https://dridk.me/expectation-maximisation.html

Regroupement hiérarchique

➜ Méthodes de classification, « ascendantes » et « descendantes »

descendante hiérarchique ➜ solution générale vers une autre plus spécifique.

➜ une seule classe contenant la totalité puis se divisent à chaque étape selon un critère jusqu'à l'obtention d'un ensemble de classes différentes.

Regroupement hiérarchique

➜ Méthodes de classification, « ascendantes » et « descendantes »

ascendantes ➜ tous les individus sont seuls dans une classe → en classes de plus en plus grandes.

➜ répartir les individus dans un certain nombre de classes.

➜ usage de similarités/distances.

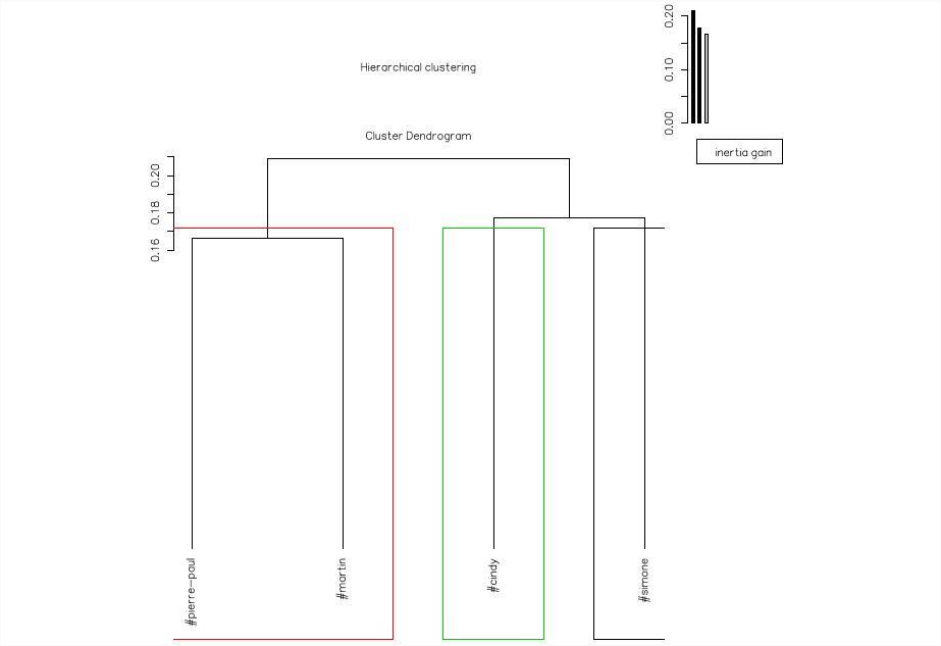

Figure 3 – `https://journals.openedition.org/revuehn/3683`, [?]

Centroïde

❋ k-moyennes.

① choix de départ : k, le nombre de classes voulues.k points au hasard parmi les n individus.

② k points = k classes ;

➜ On associe ensuite chacun des n-k points restants à la ≪ classe-point ≫ qui lui est la plus proche.

➜ chaque classe est caractérisée par la moyenne des valeurs de chacun de ses individus. On a k moyennes pour k classes.

③ La deuxième étape consiste à évaluer la distance de chaque individu à chacune des k moyennes.

➜ Certains individus peuvent ici changer de classe.

➜ A la fin de cette étape, on actualise les k moyennes.

➜ Et on réitère les étapes, jusqu'à ce qu'il y ait convergence pour obtenir nos k clusters finaux.

Centroïde

✻Limites

➜ Les classes finales dépendent beaucoup des k individus choisis pour l'initialisation.

➜ La moyenne tient parfois trop compte des valeurs aberrantes.

➜ Certains algorithmes k-means font la somme des distances des individus d'une même classe pour minimiser la variance intra-classe.
➜ D'autres représentants que le centroïde (la moyenne) ➞ le médoïde l'individu le plus central du groupe.

Densité

Zones de densité relativement élevées, zones où beaucoup de points sont proches par rapport à d'autres.

DBSCAN : Algorithme forme des classes d'individus & repère les valeurs hors du commun/bruit.

Entrées :

- ➜ la distance maximale qui peut définir deux individus comme voisins,
- ➜ le nombre minimal d'individus nécessaires pour former un groupe.

En sortie sont stockés :

- ➜ les clusters successifs
- ➜ les individus visités au fur et à mesure.

A densité

❄ Etape 1 :

✱ choisir un point parmi ceux disponibles.

✱ Les distances permettent de définir les plus proches voisins.

✗ Si le nombre minimal de point n'est pas atteint, point initial == bruit.

➥On stocke le point dans les individus visités.

✓ Si le nombre minimal de point est dépassé, point initial == initialisation d'un cluster.

➜On étudie chaque point à partir de son voisinage initial.

A densité

❄ Etape 2 :

➜ Vérification que le voisinage de chaque point comporte plus d'éléments que le minimum requis

➜ on étend le voisinage initial en le réunissant avec le voisinage du point visité.

➜ Puis on ajoute ce point dans le cluster.

A densité

❄ Etape 3 :

➜ L'opération est itérée tant que tous les individus n'ont pas été observés

➜ Lorsque tous les points du voisinage ont été testés, ceux retenus == stockage individus dans cluster

➜ A la fin on obtient :

- ➜ liste de groupes d'individus
- ➜ les individus correspondant à du bruit, ils ne sont dans aucun cluster.

Pour aller plus loin

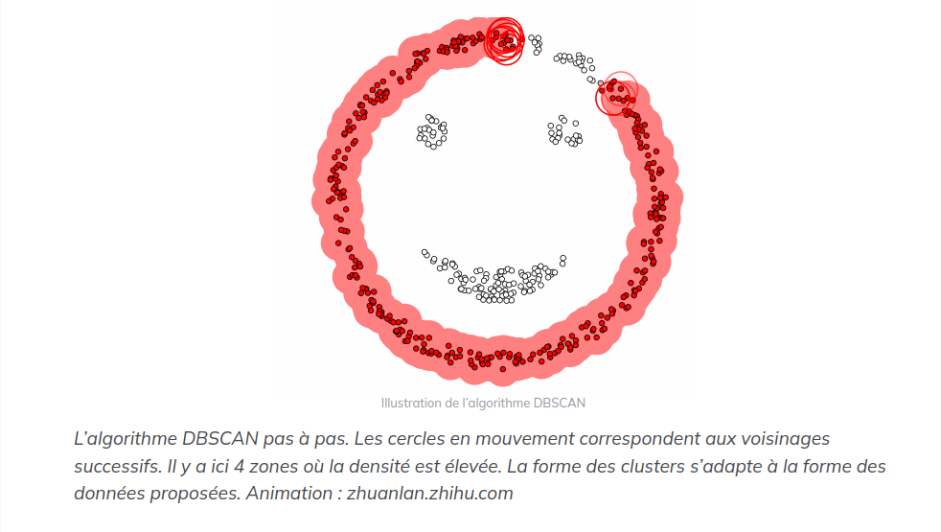

Figure 4 – `https://larevueia.fr/clustering-les-3-methodes-a-connaitre/`.

Il existe de nombreux algorithmes et méthodes pour partionner les données : `http://www.metz.supelec.fr/metz/personnel/vialle/course/BigData-2A-CS/slides-pdf/13-MachineLearning-Clustering-2spp.pdf`

References i


Ezugwu, A. E., Shukla, A. K., Agbaje, M. B., Oyelade, O. N., José-García, A., and Agushaka, J. O. (2021).
Automatic clustering algorithms : a systematic review and bibliometric analysis of relevant literature.
Neural Comput. Appl., 33(11) :6247–6306.


Melançon, J. (2023).
Analyse textométrique du lexique des personnages dans french town de michel ouellette : dire je et exprimer ses émotions.
(8).


Peng, J., Jury, E. C., Dönnes, P., and Ciurtin, C. (2021).
Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases : Applications and challenges.
Frontiers in Pharmacology, 12.

References ii