

# **Contributions aux communications multi-vues pour l'apprentissage collaboratif**

---

Denis Maurel

10 Décembre 2018

## Introduction

---

# Machine Learning

L'apprentissage machine (ou Machine Learning en anglais) est le domaine de recherche regroupant les algorithmes et les méthodes permettant **d'apprendre automatiquement** un résultat à partir d'un ensemble de données, aussi appelés **individus**. Les trois principaux sous-domaines du Machine Learning sont:

Il faut noter que les domaines ne sont pas exclusifs (apprentissage semi-supervisé, apprentissage par renforcement profond...)

L'apprentissage machine (ou Machine Learning en anglais) est le domaine de recherche regroupant les algorithmes et les méthodes permettant **d'apprendre automatiquement** un résultat à partir d'un ensemble de données, aussi appelés **individus**. Les trois principaux sous-domaines du Machine Learning sont:

- **La classification:** apprentissage des correspondances entre une donnée et son label.

Il faut noter que les domaines ne sont pas exclusifs (apprentissage semi-supervisé, apprentissage par renforcement profond...)

L'apprentissage machine (ou Machine Learning en anglais) est le domaine de recherche regroupant les algorithmes et les méthodes permettant **d'apprendre automatiquement** un résultat à partir d'un ensemble de données, aussi appelés **individus**. Les trois principaux sous-domaines du Machine Learning sont:

- **La classification:** apprentissage des correspondances entre une donnée et son label.
- **Le clustering:** détection de groupes d'individus similaires.

Il faut noter que les domaines ne sont pas exclusifs (apprentissage semi-supervisé, apprentissage par renforcement profond...)

L'apprentissage machine (ou Machine Learning en anglais) est le domaine de recherche regroupant les algorithmes et les méthodes permettant **d'apprendre automatiquement** un résultat à partir d'un ensemble de données, aussi appelés **individus**. Les trois principaux sous-domaines du Machine Learning sont:

- **La classification:** apprentissage des correspondances entre une donnée et son label.
- **Le clustering:** détection de groupes d'individus similaires.
- **L'apprentissage par renforcement:** apprentissage d'un comportement permettant à un modèle de réagir à un environnement dynamique.

Il faut noter que les domaines ne sont pas exclusifs (apprentissage semi-supervisé, apprentissage par renforcement profond...)

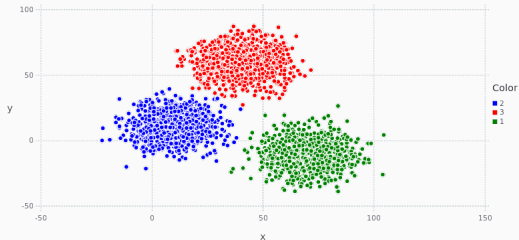
L'apprentissage machine (ou Machine Learning en anglais) est le domaine de recherche regroupant les algorithmes et les méthodes permettant **d'apprendre automatiquement** un résultat à partir d'un ensemble de données, aussi appelés **individus**. Les trois principaux sous-domaines du Machine Learning sont:

- **La classification**: apprentissage des correspondances entre une donnée et son label.
- **Le clustering**: détection de groupes d'individus similaires.
- **L'apprentissage par renforcement**: apprentissage d'un comportement permettant un modèle de réagir à un environnement dynamique.

À noter que les domaines ne sont pas exclusifs (apprentissage semi-supervisé, apprentissage par renforcement profond...)

# Clustering

- Tâche d'apprentissage **non supervisé** consistant à rassembler des groupes d'individus (a.k.a. clusters) de sorte à **maximiser la similarité intra-groupe** et à **minimiser la similarité inter-groupes**.





## Clustering: similarité entre individus

- La notion de **similarité** est souvent confondue avec la notion de **distance**.
- La similarité doit être **adaptée à la nature des données**.

Euclidienne	$\ a - b\ _2 = \sqrt{\sum_i (a_i - b_i)^2}$
Manhattan	$\ a - b\ _1 = \sum_i  a_i - b_i $
Maximum	$\ a - b\ _\infty = \max_i  a_i - b_i $
Mahalanobis	$\sqrt{(a - b)^T S^{-1} (a - b)}$
Hamming	$\text{Hamming}(a, b) = \sum_i (1 - \delta_{a_i, b_i})$

**Table 1:** Exemples de distances

## Clustering: types de partitions

Après un clustering, on obtient une partition de l'espace ainsi que les appartenances des individus à chaque groupe de cette partition. Ces appartenances peuvent être **dures**, **molles** ou **floues**.

	$C_1$	$C_2$	$C_3$
$x_1$	1	0	0
$x_2$	0	1	0
$x_3$	0	0	1
$x_4$	0	0	1

**(a)** Clustering dur

	$C_1$	$C_2$	$C_3$
$x_1$	1	1	0
$x_2$	0	1	1
$x_3$	0	0	1
$x_4$	0	0	1

**(b)** Clustering mou

	$C_1$	$C_2$	$C_3$
$x_1$	0.9	0.1	0
$x_2$	0	0.8	0.2
$x_3$	0	0.3	0.7
$x_4$	0	0	1.0

**(c)** Clustering flou

**Table 2:** Les trois principaux types d'appartenances des clusters

## Clustering: différentes approches

On peut regrouper les algorithmes de clustering en sous-catégories suivant l'approche qu'ils utilisent:

- **Méthodes hiérarchiques:** création d'un arbre de correspondance entre les individus (Agglomerative method).
- **Méthodes de quantification de vecteurs:** définition d'individus prototypes pour synthétiser les individus en entrée (K-Means).
- **Méthodes de densité:** estimation des clusters suivant les zones les plus densément peuplées de l'espace d'entrée (DBSCAN).
- **Méthodes stochastiques:** création de modèles probabilistes définissant la probabilité d'appartenance d'un individu à un cluster donné (GMM).

Apparition d'un nouveau contexte: **un même ensemble d'individu** est décrit dans plusieurs base de données indépendantes appelées **vues**.

**Problème**: comment obtenir un clustering de cet ensemble d'individus ?

Deux approches:

Apparition d'un nouveau contexte: **un même ensemble d'individu** est décrit dans plusieurs base de données indépendantes appelées **vues**.

**Problème**: comment obtenir un clustering de cet ensemble d'individus ?

Deux approches:

- **Coopérative**: chaque vue effectue un clustering de ses données locales avant de transférer ses résultats à une entité tiers qui devra fusionner les résultats.

Apparition d'un nouveau contexte: **un même ensemble d'individu** est décrit dans plusieurs base de données indépendantes appelées **vues**.

**Problème**: comment obtenir un clustering de cet ensemble d'individus ?

Deux approches:

- **Cooperative**: chaque vue effectue un clustering de ses données locales avant de transférer ses résultats à une entité tiers qui devra fusionner les résultats.
- **Collaborative**: chaque vue effectue un premier clustering local, puis le modifie en fonction des résultats obtenus par les autres vues.

Apparition d'un nouveau contexte: **un même ensemble d'individu** est décrit dans plusieurs base de données indépendantes appelées **vues**.

**Problème**: comment obtenir un clustering de cet ensemble d'individus ?

Deux approches:

- **Cooperative**: chaque vue effectue un clustering de ses données locales avant de transférer ses résultats à une entité tiers qui devra fusionner les résultats.
- **Collaborative**: chaque vue effectue un premier clustering local, puis le modifie en fonction des résultats obtenus par les autres vues.

### Exemple multi-vues

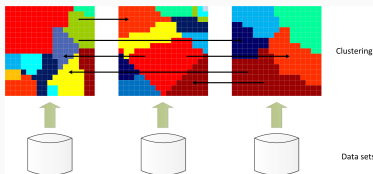
- **Vue 1:** Ensemble des achats récent d'un individus sur des sites d'e-commerce
- **Vue 2:** Salaire et régime alimentaire
- **Vue 3:** Contenu des derniers repas de chaque individu



## Clustering collaboratif: définition

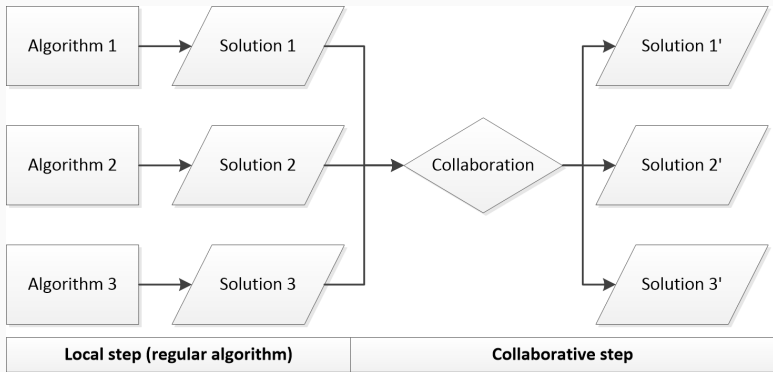
Le **clustering collaboratif** est un domaine récent désignant l'ensemble des méthodes permettant à **plusieurs algorithmes de clustering** opérant sur des **sources de données différentes** de collaborer pour **améliorer localement** leurs résultats.

- Les algorithmes utilisés peuvent être **différents**.
- Les vues doivent partager soit leurs descripteurs (clustering vertical), soit **leurs individus** (clustering horizontal) pour pouvoir être comparées.



**Figure 1:** Illustration du principe de clustering collaboratif horizontal

## Clustering collaboratif: processus



**Figure 2:** Processus de clustering collaboratif

Définition d'un algorithme de clustering collaboratif:

- $Q_{\text{local}}^i$  est généralement basé sur **le critère de l'algorithme local** à optimiser.
- $Q_{\text{collab}}^i$  se base sur **l'change d'information entre vues**, typiquement les appartenances des individus aux clusters respectifs de chaque vue.
- $\alpha_i$  et  $\beta_j^i$  sont définis **à la main**. L'approximation  $\forall j, \beta_j^i = \alpha_i^2$  est parfois utilisée car donnant de bons résultats en pratique.

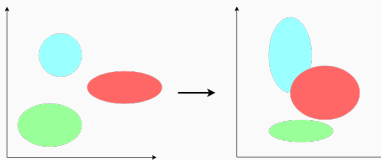
## Clustering collaboratif incrémental

---

Objectif du clustering collaboratif: définir un ensemble de clusters suivant **la distribution** des données fournies en entrée.

**Problème:** il arrive que cette distribution **évolue au cours du temps**

**Exemple:** évolution du régime alimentaire d'un individu ou de la répartition des salaires à l'échelle d'une population.



Utilisation du clustering **incrémental**: les clusters sont appris au cours du temps afin de **s'adapter** aux éventuels changements de distribution. On utilise les  $N_{\text{batch}}$  derniers individus comme échantillon d'apprentissage.

Définition d'une méthode de clustering incrémental:

Définition d'une méthode de clustering incrémental:

- Choix de la méthode de clustering

Définition d'une méthode de clustering incrémental:

- Choix de la méthode de clustering
- Adaptation de la méthode de clustering pour de l'apprentissage incrémental



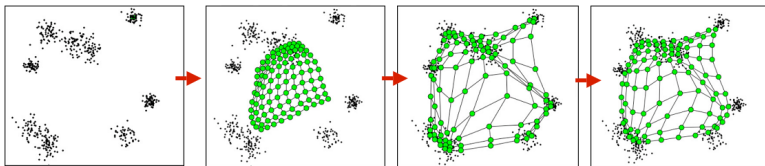
Définition d'une méthode de clustering incrémental:

- Choix de la méthode de clustering
- Adaptation de la méthode de clustering pour de l'apprentissage incrémental
- Adaptation du clustering collaboratif au modèle de clustering obtenu

## Choix de la méthode de clustering

Dans notre cas, utilisation des **cartes auto-adaptatrices de Kohonen** (ou Self-Organizing Maps (SOM) en anglais) comme méthode de clustering.

- **1<sup>re</sup> contrainte:** des individus correspondants doivent appartenir à des **prototypes correspondants** ou à leurs voisinages proches.
- **2<sup>me</sup> contrainte:** même **topologie des cartes** pour toutes les vues pour les rendre comparables.

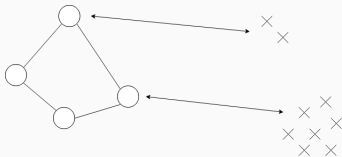


**Figure 3:** Exemple de SOM

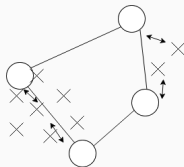
# Cartes Auto-Adaptatrices (SOM)

- Méthode basée de **prototypes** (quantification de vecteurs)
- Permet la **visualisation** de données en hautes dimensions
- Notion de **voisinage**: utilisation d'une **fonction de température**.

$$\lambda(t) = \lambda_{\min} \left( \frac{\lambda_{\max}}{\lambda_{\min}} \right)^{\frac{1}{t}} \quad K_{i,j} = \exp \left( -\frac{d_1^2(i,j)}{\lambda(t)} \right)$$



(a) Température élevée



(b) Température faible

Les SOM incrémentales ont déjà été étudiées, mais les solutions proposées se basent toutes sur **l'ajout de prototypes** (et ?).

→ **Non applicable au clustering collaboratif** du fait de la seconde contrainte: la topologie doit rester la même pour toutes les cartes.

**Limitation:** fonction de température dépendante du **temps**

**Solution:** rendre la fonction dépendante des **individus**

$$\lambda(t) = \lambda_{\min} \left( \frac{\lambda_{\max}}{\lambda_{\min}} \right)^{\frac{1}{t}} \rightarrow \tilde{\lambda}(B, W) = \frac{1}{N_{\text{batch}}} \sum_{i=1}^{N_{\text{batch}}} \|x_i - \chi(x_i)\|_2$$

L'application des SOM au clustering collaboratif se fait en définissant les termes précédemment définis:

$$Q_{\text{local}}^m = \alpha_m \sum_{i=1}^N \sum_{j=1}^{|W|} K_{j,\chi(x_i)}^m \|x_i^m - \omega_j^m\|^2$$

$$Q_{\text{collab}}^m = \sum_{m'=1, m' \neq m}^P \beta_m^{m'} \sum_{i=1}^N \sum_{j=1}^{|W|} (K_{j,\chi(x_i)}^m - K_{j,\chi(x_i)}^{m'})^2 \|x_i^m - \omega_j^m\|^2$$

- $W \rightarrow$  la carte de prototypes
- $\chi(x_i) \rightarrow$  la fonction retournant le prototype le plus proche de  $x_i$ .
- $x_i^k \rightarrow$  l'individu  $i$  dans la vue  $m$
- $\omega_j^m \rightarrow$  le prototype  $j$  de la SOM de la vue  $m$ .

Adaptation de notre version de SOM incrémentale au clustering collaboratif:

$$\lambda \rightarrow \tilde{\lambda}$$

$$K_{i,j}(\lambda) \rightarrow K_{i,j}(\tilde{\lambda}) \rightarrow \widetilde{K_{i,j}}$$

$$Q_{\text{local}}^m / Q_{\text{collab}}^m(K_{i,j}) \rightarrow Q_{\text{local}}^m / Q_{\text{collab}}^m(\widetilde{K_{i,j}}) \rightarrow \widetilde{Q_{\text{local}}^m} / \widetilde{Q_{\text{collab}}^m}$$

Le nouveau critère dépendant uniquement des  $N_{\text{batch}}$  derniers individus apparus, il est possible d'effectuer un **apprentissage collaboratif incrémental** sur l'ensemble des vues.

Les règles de mise à jour sont obtenus par **descente de gradient** appliquée sur ce critère.

- Analyse de l'impact du clustering collaboratif sur l'apprentissage incrémental.
- Application directe de la phase collaborative pour le clustering collaboratif pour pouvoir comparer les résultats aux version locales.
- Variation de la taille du batch pour étudier l'impact sur l'apprentissage.





# Système de reconstruction collaboratif

---



## Perspectives

---



## Conclusion

---



