Deep Multi-CoClustering des données multivues et multi-modales

Reda Khoufache

Dirigé par Mustapha Lebbah et Hanene Azzag

Université Paris-Saclay, UVSQ, DAVID
Université Sorbonne Paris Nord, LIPN UMR CNRS 7030

Contexte

Un volume important de données variées **multivues et multimodales** arrivent à grande vitesse.

Objectif : Développer des algorithmes capables de réaliser de façon simultanée **un multi partitionnement** des millions de données provenant de **différentes sources** et ayant différentes natures et modalités.

Multi-Coclustering et données multivues et multi-modales

Le multi-coclustering est **un algorithme Bayesien non paramétrique** qui combine deux couches:

- la première couche regroupe ensemble certaines variables dans une même vue.
- la deuxième couche qui applique un coclustering pour chaque vue.

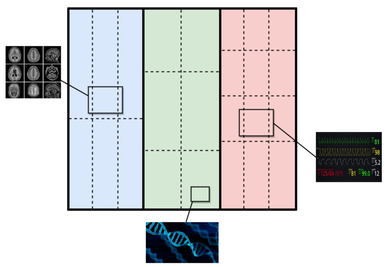


Figure 1. **Multi-Coclustering multimodal:** chaque vue regroupe des données de même nature

Définition du modèle du Multi-Coclustering:

$$x_{i,j} \mid \left\{v_j = h, w_j^h = l, z_i^h = k, \theta_{k,l}^h\right\} \sim \mathcal{N}\left(\theta_{k,l}^h\right),$$
$$\theta_{k,l}^h \sim G_0,\ v_j \sim \text{Mult}(\eta),\ w_j^h \sim \text{Mult}\left(\rho_h\right),\ z_i^h \sim \text{Mult}\left(\pi_h\right),$$
$$\eta_j(\mathbf{r}) = r_j \prod_{j'=1}^{j-1}\left(1 - r_{j'}\right), \quad r_j \overset{\text{i.i.d.}}{\sim} \text{Beta}(1, \gamma),$$
$$\rho_j^h\left(\mathbf{s}^h\right) = s_j^h \prod_{j'=1}^{j-1}\left(1 - s_{j'}^h\right), \quad s_j^h \overset{\text{i.i.d.}}{\sim} \text{Beta}\left(1, \beta_h\right),$$
$$\pi_j^h\left(\mathbf{t}^h\right) = t_j^h \prod_{j'=1}^{j-1}\left(1 - t_{j'}^h\right), \quad t_j^h \overset{\text{i.i.d.}}{\sim} \text{Beta}\left(1, \alpha_h\right),$$
$$\gamma \sim \text{Gamma}\left(a_\gamma, b_\gamma\right), \beta_h \sim \text{Gamma}\left(a_\beta, b_\beta\right), \alpha_h \sim \text{Gamma}\left(a_\alpha, b_\alpha\right).$$

Axe 1 : Multi-Coclustering topologique

La méthode **non supervisée** proposée initialement par Kohonen (carte auto-organisatrice de neurones ou **Self-Organising Map-SOM**) permet d'une manière **simultanée, un partitionnement et une discrétisation** de l'espace multi-dimensionnel et **une projection** de l'espace des données sur **un espace de faible dimension** tout en **préservant les proximités** observées dans l'espaces des données.

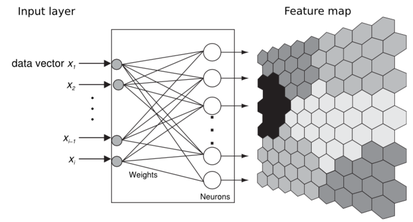
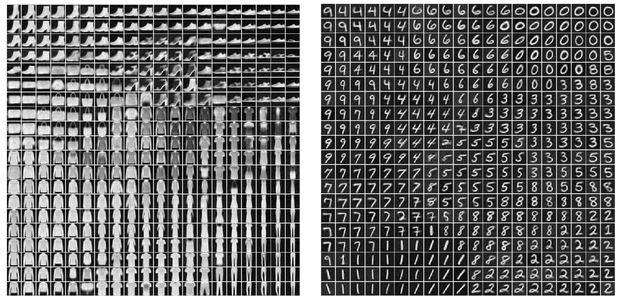


Figure 2. **Carte topologique** de Kohonen (carte **SOM**)

Cette projection doit respecter la topologie des données



(a) Projection des données Fashion MNIST sur une carte topologique

(b) Projection des données MNIST sur une carte topologique

Figure 3. **Cartes topologiques** obtenues avec une version profonde - **DESOM** (Forest et al. (2021))

Verrous

- **Extension du Multi-CoClustering** au modèle qui regroupe les variables partageant la même partition d'observations en utilisant les **modèles topologiques** à base de modèles de mélange.
- Explication des résultats du **Multi-CoClustering Topologique** et permettre un retour **visuel** plus efficace.

Axe 2 : Multi-Coclustering Scalable (distribué et fédéré)

Verrous : Développer une version **scalable** du Multi-Coclustering qui puisse être déployée sur des architectures modernes en «***clusters*** » d'ordinateurs multiples (edge computing).

Figure 4. Calcul distribué et fédéré

Axe 3 : Deep Multi-Coclustering

Verrous

- Tirer profit des réseaux de **neurones profonds** et leur capacité d'extraction des caractéristiques.
- Proposer un outil unifié qui réalise de façon **jointe l'apprentissage profond** des représentations et la tache du multi-Coclustering.

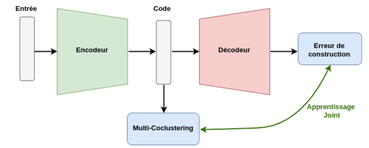


Figure 5. Architecture d'un deep multi-Coclustering basé sur les auto-encodeurs

Les premiers résultats

Développement d'un nouveau *framework* qui traite le problème de clustering d'images multivues. Le *framework* combine les *transformers* pré-entrainés avec un nouveau modèle de *multi-clustering* bayésien non paramétrique.

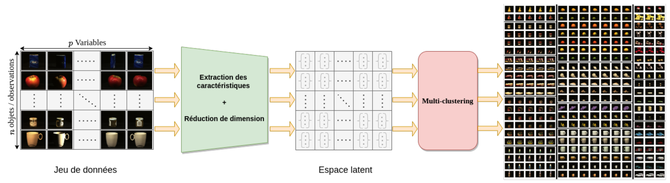


Figure 6. Le *framework* complet (Khoufache et al. (2022))

Références

Forest, F., Lebbah, M., Azzag, H., and Lacaille, J. (2021). Deep embedded self-organizing maps for joint representation learning and topology-preserving clustering. *Neural Computing and Applications*.

Goffinet, E., Lebbah, M., Azzag, H., Giraldi, L., and Coutant, A. (2022). Functional non-parametric latent block model: A multivariate time series clustering approach for autonomous driving validation. *Computational Statistics & Data Analysis*, 176(C).

Khoufache, R., Dilmi, M. D., Azzag, H., Goffinet, E., and Lebbah, M. (2022). Emerging properties from bayesian non-parametric for multiple clustering: Application for multi-view image dataset. *ICDM 2022, ORLANDO USA*.