On The Modelisation of Stability For Multi-Source Clustering

Pierre-Alexandre Murena
LTCI
Telecom ParisTech, 46 rue Barrault
Paris, France
pa.murena@telecom-paristech.fr

Denis Maurel
RDI Team
ISEP, 10 rue de Vanves
Issy-Les-Moulineaux, France
denis.maurel@isep.fr

Jérémie Sublime
RDI Team
ISEP, 10 rue de Vanves
Issy-Les-Moulineaux, France
jeremie.sublime@isep.fr

***Abstract*—TODO**

***Index Terms*—Multi-view clustering, clustering, stability, collaborative clustering**

I. Introduction

Multi-source data are an ubiquitous source of information that are produced daily and must be processed by Machine Learning algorithms. They come from multiples sources on the Internet such as social networks where data are available from several sites for the same user, but they can also be found in medical diagnosis where the combination of data from several tests can be use to better diagnose a patient, and finally satellite imaging also produces complex data where multiple types of color and texture attributes can be used to describe large images. The unsupervised exploration and processing of such data is complex process and recently gave birth to several subfields of research in Machine Learning. Multi-view clustering [1] and collaborative clustering methods [2], [3] are the two main families of algorithms that can tackle such data. Both types of methods are based on multiple clustering algorithms mining information locally in each views and then sharing them between the different algorithms. From there, the main difference between the two is that collaborative clustering only aims at sharing the information between the local algorithms with a goal of mutual improvement, while multi-view clustering shares the information and then aims at finding a single consensus clustering partition.

Many methods falling under the umbrella of multi-source clustering have been proposed in the literature for all sorts of practical applications. However, due to the relatively novelty of this field, the literature has been mostly focused on either practical aspects such as *how to exchange information between the different algorithms* -with many different algorithms proposed- [4]–[7] and *how to select which algorithms should collaborate* [8], [9] with the final goal of achieving empirical results as good as possible for the target application. As a consequences many aspects that are considered key for regular clustering algorithms have been overlooked for these methods. One of these aspect is the notion of *clustering stability* [10], a property that is considered desirable and has been widely studied in regular clustering but has never really been addressed for multi-source clustering.

Within this context, the goal of this paper are : 1) to transpose the definition of stability from regular clustering to multi-source clustering and to see which theoretical properties can be drawn from there, and 2) to assess the empirical behavior of multi-source methods in terms of this newly defined notion of stability for multi-source clustering.

The remainder of this paper is organized as follows: In section II we review the literature on clustering stability and transpose the definitions to multi-source clustering. In section III, we give some examples of empirical methods to compute stability for multi-source methods. These methods are then applied Section IV to assess the performances of several multi-source methods. Finally, this article ends with a conclusion and some perspectives on future works.

II. Stability applied to multi-source clustering

A. Definitions and notations

We start by some basic definition of what a clustering algorithm and a clustering partition are. We consider a dataspace $\mathbb{X}$ endowed with a probability measure P. If $\mathbb{X}$ happens to be a metric space, we denote l its metric. A sample $S = \{x_1, \cdots, x_m\}$ is drawn i.i.d. from $(\mathbb{X}, P)$.

Within this context, a clustering algorithm $\mathcal{A}$ is a function $\mathcal{A} : X \to \mathcal{C}$ which from any finite sample $S \subset X$ -($X \subset \mathbb{X}$)- creates a clustering C. A clustering C is also a function $C : X \to \mathbb{N}$ which to any data subset subset $X \subset \mathbb{X}$ associates a solution vector in the form of matching clusters. The clusters are defined by $C_i = C^{-1}(\{i\}) = \{x \in X; C(x) = i\}$. With this definition, a clustering is a partitioning of the entire data space, which when applied to a specific sample of the dataset gives the matching clusters.

Definition 1: Clustering distance Let $\mathcal{P}$ be a family of probability distribution over some domain $\mathbb{X}$. Let Σ be a family of clusterings of $\mathbb{X}$. A clustering distance is a function $d : \mathcal{P} \times \Sigma \times \Sigma \to [0, 1]$ satisfying for any $P in \mathcal{P}$ and any $C_1, C_2, C_3 \in \Sigma$:

1) $d_P(C_1, C_1) = 0$
2) $d_P(C_1, C_2) = d_P(C_2, C_1)$ (symmetry)
3) $d_P(C_1, C_3) \leq d_P(C_1, C_2) + d_P(C_2, C_3)$ (triangle inequality)

Note that d_P is not properly a distance, but rather a metric, since we do not require $d_P(C_1, C_2) = 0 \Rightarrow C_1 = C_2$.

Clustering stability measures how a perturbation in the data will affect the result of a clustering algorithm. In [10], the stability of an algorithm $\mathcal{A}$ for a sample of size m w.r.t. a probability distribution P is defined as follows:

Definition 2: Clustering stability Let P be the probability distribution over $\mathcal{X}$, d be a clustering distance and $\mathcal{A}$ a clustering algorithm. The stability of $\mathcal{A}$ for the sample size m w.r.t. the probability distribution P is:

$$stab(\mathcal{A}, P, m) = \mathop{\mathbb{E}}_{\substack{X_1 \in P^m \\ X_2 \in P^m}} [d_P(\mathcal{A}(X_1), \mathcal{A}(X_2))] \quad (1)$$

Then, the stability of $\mathcal{A}$ w.r.t to P is:

$$stab(\mathcal{A}, P) = \limsup_{m \to \infty} stab(\mathcal{A}, P, m) \quad (2)$$

And an algorithm is said to be stable when $stab(\mathcal{A}, P) = 0$.

B. Transposition to multi-source clustering

In the context of multi-source clustering, we consider that the total dataspace $\mathbb{X}$ can be decomposed into the product $\mathbb{X}^1 \times \cdots \times \mathbb{X}^J$ of J view spaces $\mathbb{X}^j$. From there, based on the description of multi-source clustering from [2], we can get the following definition:

Definition 3: Global clustering A global clustering is defined as a combination of local clustering in the following sense: A global clustering C of the subset $\mathcal{X} \subset \mathbb{X}$ is a function $C : \mathcal{X} \to \mathbb{N}^J$. The i-th cluster for view j, denoted C_i^j is defined as:

$$C_i^j = \{x \in \mathcal{X}; (C(x))^j = i\} \subset \mathbb{X} \quad (3)$$

Likewise, a global collaborative algorithm $\mathcal{A} = \langle \mathcal{A}^1, \cdots, \mathcal{A}^J \rangle$ is a function which compute such global clustering based on local clustering partitions C^j on $\mathbb{X}^j$. More formally, if we denote by $\mathbb{A}^j$ the set of clustering algorithms on $\mathbb{X}^j$, C the set of global clustering on $\mathcal{X} \subset \mathbb{X}$ and Σ the set of finite partitions of $\mathcal{X}$, then a multi-source clustering algorithm is a mapping $\mathcal{A}^1 \times \cdots \times \mathcal{A}^J \times \sigma \to C$.

In general, the projection of the clustering obtained by a multi-source algorithm into one of the view j is distinct of what would have been obtain by the local algorithm $\mathcal{A}^j$ alone: If $C = \mathcal{A}(X)$, then in general $C^j \neq \mathcal{A}^j(X^j)$. It is this property that makes multi-source clustering useful, because it ensure that multi-source collaboration makes it possible solutions that could not have been found by local algorithms alone.

- **balancer Proposition 13 et Lemme 1 de la these de PAM**

III. Empirical assessment of the stability

In the previous section, we have seen that stability with multi-source clustering is defined in the same way than for regular clustering up to an isomorphism. However, while the work of Shai Ben David et al. [10] provides a few examples of concrete clustering distance for regular stability, there are no known examples of distances between multi-source clusterings.

In this section, we propose the second contribution of this paper in the form of examples of clustering distances that can be used to empirically compute the stability of multi-source clustering algorithms. Our proposal is in no way exhaustive, and other methods are certainly possible.

The main difficulty when it comes to comparing and computing the distance between two multi-source partitions is that we are not computing the distance between two local clustering partitions, but between two sets of local clustering partitions. To do so, we propose two main solutions that are detailed in the two next subsections.

A. Mean distance partition to partition

The simplest distance that we can think of between two multi-view partitions that use the same views is the following: computing the mean distance between each views pairwise. Let $d^j_{P^j}$ be a local canonical distance for view j, then the global distance between two partitions C_1 and C_2 can be defined as follows:

$$d_P(C_1, C_2) = \frac{1}{J} \sum_{j=1}^{J} d^j_{P^j} \left(C_1^j, C_2^j \right) \quad (4)$$

Note that this distance assumes that the view to view pairwise mapping is the best mapping solution.

- **Proposition 12 de PAM**

B. Gromov-Hausdorff distance between multi-source partitions

Our second proposal for a global clustering distance comes from a weakness of the distance defined in Equation (5): it relies solely on the difference of partitions between equivalent views C_1^j and C_2^j and never consider the structural differences between the two global clustering.

To take that into consideration, we propose to adapt the technique used in [11], where the authors tackle the issue of stability in hierarchical clustering. In our case, the distance between two global partitions would be computed using the following steps:

1) Map the local clusterings from C_1 and C_2 using any relevant mapping method. For example we can use the same pairwise direct mapping as in the previous distance an map each $(C_1)^j$ and $(C_2)^j$ together since they are the same views.
2) Build the two dendrogram for C_1 and C_2 using regular clustering distances.
3) Use the dendrograms to compute the GromovHausdorff distance between the two multi-source partitions as proposed in [11].

- **Lire Carlsson et mettre des figures pour expliquer ici.**
- **Penser au cas merdique ou les $(C_1)^j$ avec $(C_2)^j$ ne serait pas le mapping le plus pertinent**
- **Choisir une distance pour construire les dendrogrammes**

- **Figure pour Gromov-Hausdorff sur dendrogramme et pas sur que ce soit direct (lire Carlsson)**

C. Picking the right clustering distance function

As one can see, with either of the two previously proposed distances, it all comes down to picking a pairwise regular clustering distance (a local $d^j_{P^j}$) to do the actual computation.

A few of them are proposed for prototype based clustering algorithm in [10]. However, in the general case these distance cannot be exactly computed due to the impossibility of formulating the clustering function. As such, in practice the distance functions are approached by using the partition vectors assigned to given samples of the dataset [12].

Computing distance between partition vectors of the same sample can be easily done based on elements such as the confusion entropy [4], [13] or even simple pairwise vector comparisons such as the Rand Index, or cosine similarity. From there, bootstrap technique or large sampling can be used to approach the clustering distance.

$$d_P(C_1, C_2) = \lim_{m\to\infty} \int_{X\in P^m} d\left(C_1(X), C_2(X)\right) dX \qquad (5)$$

IV. Experiments

Dire qu'on va utiliser [14]

A. Raw stability using bootstrap

- **Echantillons et nombre de vues fixe, initialisation non-fixe, on lance de multiple fois le młme modle collaboratif et on calcule la distance moyenne entre les diffrentes partititions qui sortent. A tester sur plusieurs algos. Sortir un tableau de chiffres**
- **Faire avec les 2 distances**

B. Influence of the sample size

- **Nombre de vue fixe, initialisation fixe, on fait monter la taille de l'chantillon et on regarde comment la stabilit volue. Faire des courbes d'volutions pour plusieurs types d'algo : K-Means, GMM, DBSCAN, etc.**
- **Faire avec les 2 distances**

C. Influence of the number of views

- **Taille de l'chantillon fixe, initialisation fixe, on fait monter le nombre de vues et on regarde comment la stabilit volue. Faire des courbes d'volutions pour plusieurs types d'algo : K-Means, GMM, DBSCAN, etc. Attention bien normaliser par le nombre de vues.**
- **Faire avec les 2 distances**

V. Conclusion and future works

- **Todo**

References

[1] A. Zimek and J. Vreeken, "The blind men and the elephant: on meeting the problem of multiple truths in data from clustering and pattern mining perspectives," *Machine Learning*, vol. 98, no. 1-2, pp. 121–155, 2015. [Online]. Available: http://dx.doi.org/10.1007/s10994-013-5334-y

[2] A. Cornuéjols, C. Wemmert, P. Gançarski, and Y. Bennani, "Collaborative clustering: Why, when, what and how," *Information Fusion*, vol. 39, pp. 81–95, 2018.

[3] P. Vanhaesebrouck, A. Bellet, and M. Tommasi, "Decentralized Collaborative Learning of Personalized Models over Networks," in *AISTATS*, 2017.

[4] J. Sublime, B. Matei, G. Cabanes, N. Grozavu, Y. Bennani, and A. Cornuéjols, "Entropy Based Probabilistic Collaborative Clustering," *Pattern Recognition*, vol. 72, pp. 144–157, 2017.

[5] M. Ghassany, N. Grozavu, and Y. Bennani, "Collaborative clustering using prototype-based techniques," *International Journal of Computational Intelligence and Applications*, vol. 11, no. 3, 2012.

[6] G. Cleuziou, M. Exbrayat, L. Martin, and J. Sublemontier, "Cofkm: A centralized method for multiple-view clustering," in *ICDM 2009, The Ninth IEEE International Conference on Data Mining, Miami, Florida, USA, 6-9 December 2009*, W. Wang, H. Kargupta, S. Ranka, P. S. Yu, and X. Wu, Eds. IEEE Computer Society, 2009, pp. 752–757. [Online]. Available: http://dx.doi.org/10.1109/ICDM.2009.138

[7] C. Wemmert and P. Gancarski, "A multi-view voting method to combine unsupervised classifications," *Artificial Intelligence and Applications, Malaga, Spain,*, pp. 447 – 452, 2002.

[8] J. Sublime, B. Matei, and P.-A. Murena, "Analysis of the influence of diversity in collaborative and multi-view clustering," in *2017 International Joint Conference on Neural Networks, IJCNN 2017*, 2017.

[9] P. Rastin, G. Cabanes, N. Grozavu, and Y. Bennani, "Collaborative clustering: How to select the optimal collaborators?" in *IEEE Symposium Series on Computational Intelligence, SSCI 2015, Cape Town, South Africa, December 7-10, 2015*. IEEE, 2015, pp. 787–794. [Online]. Available: http://dx.doi.org/10.1109/SSCI.2015.117

[10] S. Ben-David, U. Von Luxburg, and D. Pál, "A sober look at clustering stability," in *International Conference on Computational Learning Theory*. Springer, 2006, pp. 5–19.

[11] G. Carlsson and F. Mémoli, "Characterization, stability and convergence of hierarchical clustering methods," *J. Mach. Learn. Res.*, vol. 11, pp. 1425–1470, Aug. 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=1756006.1859898

[12] U. von Luxburg, "Clustering stability: An overview," *Foundations and Trends in Machine Learning*, vol. 2, no. 3, pp. 235–274, Mar. 2010.

[13] X.-N. Wang, J.-M. Wei, H. Jin, G. Yu, and H.-W. Zhang, "Probabilistic confusion entropy for evaluating classifiers," *Entropy*, vol. 15, no. 11, pp. 4969–4992, 2013.

[14] P. Murena, J. Sublime, B. Matei, and A. Cornuéjols, "An information theory based approach to multisource clustering," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, J. Lang, Ed. ijcai.org, 2018, pp. 2581–2587. [Online]. Available: https://doi.org/10.24963/ijcai.2018/358