Cartes Auto-Organisatrices Incrémentales appliquées au Clustering Collaboratif

Denis Maurel[*,**], Jérémie Sublime[*]
Sylvain Lefebvre[*]

[*]LISITE, ISEP, 28 rue Notre Dame des Champs, 75006 Paris France
prenom.nom@isep.fr,
[**]CEDRIC, CNAM, 292 rue Saint-Marin 75003 Paris FRANCE

Résumé. Le Clustering Collaboratif (CC) vise à faire ressortir les structures communes présentes dans plusieurs vues indépendantes en se basant sur une première étape de clustering locale, effectuée dans notre cas à l'aide de Cartes Auto-Organisatrices (SOM pour Self Organizing Maps en anglais). Pour faire face à la quantité toujours croissante de données disponibles, l'utilisation de méthodes de clustering incrémentales est devenue nécessaire. Ce papier présente un algorithme de SOM incrémentales compatibles avec les contraintes du CC. Les expérimentations conduites sur plusieurs jeux de données démontrent la validité de cette méthode et présentent l'influence de la taille du batch utilisé lors de l'apprentissage.

1 Introduction

Dans cet article, nous étudions le clustering conjoint de plusieurs bases de données distribuées (nommées vues), aussi appelé Clustering Collaboratif (CC). Une méthode de CC applicable à des données arrivant en continu permet de résoudre des problèmes en temps réel avec une contrainte de confidentialité sur les données. Cet article constitue un résumé de l'article publié dans la conférence ICONIP 2017 (Maurel et al. (2017)).

L'objectif du CC (Cornuéjols et al. (2018)), est de trouver une manière de partitionner un même ensemble d'individus décrits par différents ensembles de caractéristiques. Pour ce faire, les vues vont échanger des informations, sans pour autant échanger les valeurs qu'elles contiennent (dans un soucis de confidentialité). Cet objectif est atteint par l'intermédiaire de vecteurs synthétisant l'information contenue dans chaque base sous forme d'individus représentatifs de la distribution des données. Ces individus sont appelés les prototypes de la vue.

Le CC peut être décomposé en deux phases : la phase locale durant laquelle chaque algorithme de clustering est appliqué localement afin d'obtenir les prototypes de la vue, et la phase collaborative, durant laquelle chaque vue fournit à ses pairs les informations sur ses prototypes afin de partager ce qui a été appris localement.

La création d'une méthode de CC incrémental présente plusieurs défis. Le premier est que, dans notre cas, le CC se base sur des algorithmes de clustering à base de prototypes (Ghassany et al. (2013)), réduisant de fait le nombre de méthodes utilisables. Le deuxième est

que les méthodes de clustering incrémentales ne sont pas forcément compatibles avec le CC, même si elles se basent sur des prototypes. Enfin, le troisième défi se trouve dans l'adaptation nécessairement *ad hoc* des règles de mise à jour collaboratives suivant l'algorithme employé.

Ce papier présente une méthode d'apprentissage de SOM incrémentales robuste aux éventuelles évolutions de la distribution des données et compatible avec le paradigme du CC. La composante principale de cette approche se trouve dans la modification de la fonction de température des SOM, qui devient indépendante du temps. À notre connaissance, ce type de méthode n'a encore jamais été proposé dans la littérature.

Cet article est organisé comme suit : un bref bilan des méthodes de clusterings incrémentales et collaboratives est présentée dans la Section 2. Notre approche sur les SOM incrémentales et sur leurs applications au CC est présentée dans la Section 3, suivie par les résultats expérimentaux présentés en Section 4. Une conclusion ainsi qu'une ouverture sur les futures pistes à suivre sont présentées en Section 5.

2 Recherches associées

Un état de l'art sur le CC peut être trouvé dans Cornuéjols et al. (2018). Cet article présente les principales spécificités du domaine ainsi que ses principaux défis.

Le CC basé sur les SOM a été étudié dans Grozavu et al. (2014); Rastin et al. (2015) de même que sa version basée sur les Generative Topographic Mapping (GTM) (Sublime et al. (2017); Ghassany et al. (2013)) et sur Fuzzy C-Means (Pedrycz et Rai (2008); Mitra et al. (2006)). Néanmoins, les méthodes proposées dans ces articles ne fonctionnent pas dans le contexte incrémental qui est étudié ici. Cette contrainte a déjà été étudiée pour les SOM non-collaboratives dans des travaux uniquement dédiés au clustering incrémental : les méthodes proposées par Deng et Kasabov ou encore Papliński (2012) reposent ainsi sur l'évolution topologique de la SOM au cours du temps. Ces références mettent en lumière que même si des méthodes existent pour chaque sous-partie du problème, il n'existe à notre connaissance aucune méthode permettant de faire du CC basé sur les SOM sans modification topologique de ces dernières.

3 CC incrémental basée sur les SOM

3.1 CC Incrémental

La principale limitation du CC basé sur les SOM est à l'heure actuelle que les SOM incrémentales sont toutes basées sur la modification topologique de leurs cartes de neurones. Ce genre de modification n'est pas permis par les règles de mise à jour du CC. En effet, ce paradigme suppose que chaque vue décrit ses individus en utilisant le même nombre de prototypes, et ce pour deux raisons : permettre la comparaison entre vues et garder les correspondances topologiques entre chaque paire de vues. Dans notre cas, les prototypes correspondent aux neurones des SOM, et de fait, la topologie de l'ensemble des SOM doit être unique et fixée au lancement de l'algorithme. De plus, le clustering incrémental se doit d'être réactif aux éventuelles modifications de la distribution des données au cours du temps.

3.2 SOM incrémentale

Dans notre version incrémentale des SOM, nous considérons que les données arrivent en continu. Ainsi, nous supposons qu'à chaque instant, le modèle n'a connaissance que du batch B des N_{batch} derniers individus. La contrainte incrémentale ne permettant pas de définir de temps d'arrêt, notre méthode présente une variation de la fonction de température afin d'éviter toute dépendance temporelle. La nouvelle fonction de température $\widetilde{\lambda}$ est définie par :

$$\widetilde{\lambda}(B, W) = \frac{1}{N_{batch}} \sum_{i=1}^{N_{batch}} \|x_i - \chi(x_i)\|_2 \tag{1}$$

Où x_i et $\chi(x_i)$ correspondent respectivement à un individu du batch et à son plus proche prototype dans la vue étudiée. Cette fonction $\widetilde{\lambda}$ est ensuite bornée entre λ_{min} et λ_{max} afin d'éviter qu'elle ne prenne des valeurs extrêmes, ce qui entraînerait des modifications trop importantes de la topologie de la carte.

Cette définition permet à la carte d'être réactive à la nouveauté. Si les éléments d'un batch sont éloignés des neurones courants, l'ensemble de la carte aura besoin d'être ajusté (hautes valeurs de $\widetilde{\lambda}$). À l'inverse, si les individus sont proches des prototypes courants, la carte n'aura besoin que de d'ajustements locaux (faibles valeurs de $\widetilde{\lambda}$). La fonction de voisinage définie par $\widetilde{\lambda}$ sera désignée par $\widetilde{K}$. Dans la suite de cet article, $\widetilde{K}^m_{i,j}$ désignera la valeur de la fonction de voisinage entre les neurones i et j de la m-ème SOM (et donc de la m-ème vue).

3.3 Adaptation au CC

On considère les bases de données $\{X[i] |\ i \in 1..P\}$ contenant le même ensemble d'individus décrits par différents ensembles de caractéristiques, avec P modèles (ici des SOM) entraînés à représenter chacune des vues. Afin de clarifier les notations, $W^{m \in \{1..P\}}$ désignera le m-ème modèle créé en utilisant la m-ème base de donnée. On impose de plus le critère d'apprentissage suivant : des neurones correspondants ou leurs voisinages proches devront capturer les même individus indépendamment de la vue considérée.

Afin d'adapter le critère original à la version incrémentale du CC, nous l'approximons en utilisant la nouvelle fonction de voisinage $\widetilde{K}$ et en sommant les distances sur le batch courant plutôt que sur l'intégralité des individus :

$$\widetilde{R}^m(\chi, \omega) = \widetilde{R}_{Local}(W) + \widetilde{R}_{Collab}(W) \tag{2a}$$

$$\widetilde{R}_{Local}(W) = \alpha_m \sum_{i=1}^{N_{batch}} \sum_{j=1}^{|W|} \widetilde{K}^m_{j,\chi(x_i)} \|x_i^k - \omega_j^k\|^2 \tag{2b}$$

$$\widetilde{R}_{Collab} = \sum_{m'=1, m' \neq m}^{P} \beta_m^{m'} \sum_{i=1}^{N_{batch}} \sum_{j=1}^{|W|} (\widetilde{K}^m_{j,\chi(x_i)} - \widetilde{K}^{m'}_{j,\chi(x_i)})^2 \|x_i^m - \omega_j^m\|^2 \tag{2c}$$

Avec α et β les coefficients de collaboration fixés et qui définissent les pondérations des termes locaux et collaboratifs dans le critère d'apprentissage. Une synthèse du CC horizontal incrémental peut être trouvée dans l'Alg. 1. Dans un soucis de concision, la formule de

Algorithm 1 CC horizontal incrémental

1. Initialisation
$\forall m \in 1..P$, $W^m \leftarrow$ Initialisation du m-ème modèle de la m-ème vue
2. Apprentissage incrémental
loop
  if Arrivée d'un individu **then**
    Mise à jour du batch comme une file d'individus (premier arrivé premier sorti)
    2.1. Étape locale
    $\forall m \in 1..P$, Mise à jour des prototypes de W^m (SOM incrémentale)
    2.2. Étape collaborative
    for $m \in 1..P, \omega \in W^m$ **do**
      $\omega = \omega + (\Delta\omega)_{collab}$
    end for
  end if
end loop

$(\Delta\omega)_{collab}$ n'a pas été explicitée ici. Elle peut être trouvée dans Ghassany et al. (2013) et est obtenue est dérivant le critère de l'Eq. 2 par rapport au point ω à mettre à jour.

4 Résultats Expérimentaux

4.1 Bases de données et mesures de qualité

Afin d'évaluer la méthode présentée dans ce papier, nous l'avons testée sur quatre bases de données : Spam Base, Waveform, Wisconsin Diagnostic Breast Cancer (WDBC) et Isolet [1].

Durant ces expérimentations, chaque base a été normalisée puis divisée en 3 vues contenant chacune un tiers des variables originales. Nous supposons ici que l'on dispose de suffisamment d'information sur chaque variable pour permettre sa normalisation au moment où elle apparaît, par exemple en connaissant ses bornes. Les mesures de qualité ici utilisées sont l'erreur de quantification et l'index de pureté communément utilisés pour l'analyse de SOM.

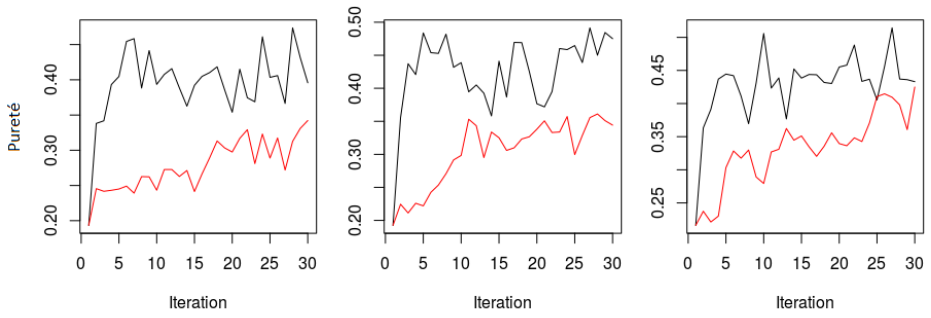
4.2 Expérimentations

Dans un soucis de concision, seuls les résultats obtenus sur Isolet sont présentés, ces derniers pouvant être généralisés pour l'ensemble des bases étudiées. Les SOM utilisées sont composées de 10×10 neurones, avec $\lambda_{min} = 0.3$, $\lambda_{max} = 3$, $\epsilon = 0.5$ (pas d'apprentissage fixé durant nos expérimentations), $N_{batch} = 10$. Ces paramètres ont été obtenus empiriquement après plusieurs apprentissages.

Les puretés respectives des cartes peuvent être trouvées sur les Fig. 1a, Fig. 1b et Fig. 1c. Il apparaît que le CC améliore la pureté au détriment de la stabilité par rapport aux SOM incrémentales. La stabilité réfère ici à l'écart type de la pureté au cours du temps. Cette instabilité peut être causée par l'apprentissage par batchs qui, par définition, ne sont pas représentatifs de

1. http ://archive.ics.uci.edu/ml/index.php

la population globale. Néanmoins, le fait que la méthode aboutisse quand même à un résultat laisse penser que les biais successifs se compensent sur le long terme.



(a) $N_{batch} = 10$, Vue 1 (b) $N_{batch} = 10$, Vue 2 (c) $N_{batch} = 10$, Vue 3

FIG. 1: Évolution des puretés pour la base Isolet. Les lignes rouges représentent les SOM incrémentales tandis que les lignes noires représentent les SOM collaboratives. Chaque itération correspond à l'utilisation d'un nouveau batch.

Dans un second temps, l'impact de la phase collaborative de notre méthode sur le clustering final a été étudié. Sur chaque base de données, deux apprentissages ont été effectué : le premier intitulé SOM Incrémentale (SOMI) ne comprenait que la phase locale de l'Alg. 1 tandis que le second, intitulé CC Incrémental (CCI), comprenait l'ensemble de la méthode. Les résultats de cette expériences sont présentés dans Tab. 1.

	Spam Base			Waveform			WDBC			Isolet		
	1	2	3	1	2	3	1	2	3	1	2	3
SOMI	0.31	**0.18**	0.18	**0.18**	**0.17**	**0.24**	**0.19**	**0.16**	0.20	2.15	2.84	2.85
CCI	**0.26**	0.19	**0.16**	0.23	0.19	0.30	0.19	0.19	**0.16**	**1.27**	**1.38**	**1.37**

TAB. 1: Erreur de quantification sur chaque base de données. Les nombres en gras sont les plus petits de chaque colonne.

Les résultats présentés nous permettent de conclure que dans la plupart des cas, la phase collaborative n'améliore pas significativement les résultats des SOM incrémentales utilisées seules. Cependant dans le cas de la base Isolet, la phase collaborative améliore nettement les résultats obtenus. Il est apparu que la base Isolet était une base *clairsemé* (ou *sparse* en anglais). Cette caractéristique couplée à la division de la base initiale en 3 vue a réduire l'information déjà limitée disponible pour chaque clustering local. Avec l'ajout de la phase collaborative, l'information contenue dans chaque vue a pu être partagée, ce qui a permis de nettement améliorer les résultats par rapport aux clustering locaux. C'est ce résultat qui nous permet de justifier l'intérêt de l'ajout de la phase collaborative de notre méthode.

5 Conclusion et Futures Recherches

Dans cette étude, nous avons présenté une méthode permettant de générer des SOM incrémentales sans modifications topologiques ainsi que leurs application au CC horizontal. Cette méthode se base sur une nouvelle fonction de température $\widetilde{\lambda}$ qui ne dépend plus que des données arrivantes, ce qui permet l'adaptation de la carte à un flux de données continu. Les méthodes présentées ont été testées sur 4 bases de données différentes. L'influence du paramètre N_{batch} sur la stabilité de l'apprentissage a aussi été analysée.

Pour poursuivre ces travaux, nous prévoyons d'adapter notre méthode aux GTM, par nature proche des SOM et qui sont susceptibles d'améliorer la qualité de l'apprentissage.

Références

Cornuéjols, A., C. Wemmert, P. Gançarski, et Y. Bennani (2018). Collaborative clustering : Why, when, what and how. *Information Fusion 39*, 81–95.

Deng, D. et N. Kasabov. Esom : An algorithm to evolve self-organizing maps from online data streams. In *Neural Networks, 2000*, Volume 6, pp. 3–8. IEEE.

Ghassany, M., N. Grozavu, et Y. Bennani (2013). Collaborative multi-view clustering. In *The 2013 International Joint Conference on*, pp. 1–8. IEEE.

Grozavu, N., G. Cabanes, et Y. Bennani (2014). Diversity analysis in collaborative clustering. In *2014 International Joint Conference on*, pp. 1754–1761. IEEE.

Maurel, D., J. Sublime, et S. Lefebvre (2017). Incremental self-organizing maps for collaborative clustering. Springer.

Mitra, S., H. Banka, et W. Pedrycz (2006). Rough–fuzzy collaborative clustering. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 36*(4), 795–805.

Papliński, A. P. (2012). Incremental self-organizing map (isom) in categorization of visual objects. In *ICONIP*, pp. 125–132. Springer.

Pedrycz, W. et P. Rai (2008). Collaborative clustering with the use of fuzzy c-means and its quantification. *Fuzzy Sets and Systems 159*(18), 2399–2427.

Rastin, P., G. Cabanes, N. Grozavu, et Y. Bennani (2015). Collaborative clustering : How to select the optimal collaborators ? In *Computational Intelligence, 2015 IEEE*, pp. 787–794.

Sublime, J., B. Matei, G. Cabanes, N. Grozavu, Y. Bennani, et A. Cornuéjols (2017). Entropy based probabilistic collaborative clustering. *Pattern Recognition 72*, 144–157.

Summary

Collaborative clustering aims at revealing the common structures of data distributed on different sites using local clustering methods such as Self-Organizing Maps (SOM). To face the ever growing quantity of data available, incremental clustering methods are required. This paper presents an algorithm to perform incremental SOM-based collaborative clustering. The experiments conducted on several datasets demonstrate the validity of the method and present the influence of the batch size on the learning.