Vers un cadre évolutif de classification non supervisée

Mohamed Charouel*, Minyar Sassi-Hidri** Mohamed Ali Zoghlami***

Université Tunis El Manar
Ecole Nationale d'Ingénieurs de Tunis
BP. 37 le Belvédère 1002 Tunis, Tunisie
{*mohamed.charouel, **minyar.sassi}@enit.rnu.tn, ***ma.zoghlami@gmail.com

Résumé. La classification non supervisée (clustering) évolutive surpasse généralement par celle statique en produisant des groupes de données (clusters) qui reflètent les tendances à long terme tout en étant robuste aux variations à court terme. Dans ce travail, nous présentons un cadre différent pour le clustering évolutif d'une manière incrémentale par un suivi précis des variables de proximité temporelles entre les objets suivis par un clustering statique ordinaire.

1 Introduction

Dans de nombreuses applications pratiques de gestion de clusters, résultat d'une opération de classification non supervisée (clustering), les objets à classifier évoluent dans le temps. Le but est donc d'obtenir des clusters optimaux à chaque pas de temps (Falkowski et al. (2006)).

Tang et al. (2008) et Zhang et al. (2009) ont proposés des méthodes de clustering évolutif dans le but de produire des clusters optimaux qui reflètent les dérives à long terme dans les objets tout en étant robuste aux variations à court terme. Chi et al. (2009) ont développé cette idée en proposant deux cadres évolutifs pour le clustering spectral : *PCQ* (Preserving Cluster Quality) et *PCM* (Preserving Cluster Membership). Les deux cadres ont été proposés afin d'optimiser la modification de la fonction de coût proposée initialement par Chakrabarti et al. (2006).

Notre travail adapte le principe d'incrémentabilité afin de le généraliser à un ensemble d'algorithmes de clustering. Le cadre proposé consiste à estimer les états de données à l'aide des proximités qui sont à la fois actuels et passés. Puis, il effectue un clustering statique sur les estimations de ces états. Ce cadre de suivi de clustering évolutif d'une manière incrémentale a été utilisé pour étendre une variété d'algorithmes de clustering statiques tels que les C-Moyennes Floues (*CMF*) de Bezdek (1984), les *k-moyennes* de Mac-Queen (1967) et les approches spectrales de clustering présentées par Filippone et al. (2008).

Le reste de ce papier est organisé comme suit : en section 2, nous présentons le cadre évolutif du clustering. La section 3 présente les résultats d'expérimentation du cadre proposé sur une variété différente d'algorithmes de clustering statiques. La section 4 conclue le papier et présente les travaux futures.

2 Cadre de clustering évolutif

Nous traitons le clustering évolutif comme étant un problème de suivi par un regroupement statique ordinaire. Pour ceci, nous étudions des matrices de proximité, notées W^t, comme la réalisation d'un processus aléatoire non stationnaire indexé par des mesures de temps discrètes. Elles sont données par l'équation (1).

$$W^t = \Psi^t + N^t, \quad t = 0, 1, 2, ... \tag{1}$$

Où W^t est une matrice déterministe inconnue des états non observés et N^t est une matrice de bruit de moyenne nulle. Ψ^t change au fil du temps pour réfléchir à long terme des dérives dans les proximités. Nous présentons une approche plus simple qui implique une mise à jour recursive des estimations de ces états en utilisant un seul paramètre α nommé facteur d'oubli.

2.1 Estimation de la matrice de proximité

Une meilleure estimation peut être obtenue en utilisant un lissage de la matrice de proximité $\overline{W^t}$ définie dans l'équation (2).

$$\overline{W^t} = \alpha^t \overline{W^{t-1}} + (1 - \alpha^t) W^t \; pour \;\; t \geq 1 \; et \;\; \overline{W^0} = W^0 \tag{2}$$

Cette matrice lissée est un candidat dans l'estimation de Ψ^t. Des méthodes d'estimation ont été proposées dans Ledoit et Wolf (2003), Schäfer et Strimmer (2005) et Chen et al. (2010). D'une manière générale, ils calculent la différence entre la matrice de proximité réelle et lissée donnée par l'équation (3).

$$L(\alpha^t) = \| \; \overline{W^{t-1}} - \Psi^t \; \|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} (\overline{W_{ij}^t} - \Psi_{ij}^t)^2 \tag{3}$$

Puisque $N^t, N^{t-1}, ..., N^0$ sont mutuellement indépendants et ont une moyenne nulle et la variance conditionnelle de $\overline{W^{t-1}}$ est nulle, le risque de l'espérance conditionnelle de la fonction de perte peut être alors exprimé dans l'équation (4).

$$R(\alpha^t) = \sum_{i=1}^{n} \sum_{j=1}^{n} \{(1 - \alpha^t)^2 Var(n_{ij}^t) + (\alpha^t)^2 (\overline{W^{t-1}} - \Psi_{ij}^t)^2\} \tag{4}$$

La dérivée première correspondante au facteur d'oubli est donnée par l'équation (5).

$$(\alpha^t)^* = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} Var(n_{ij}^t)}{\sum_{i=1}^{n} \sum_{j=1}^{n} \{(\overline{W^{t-1}} - \Psi_{ij}^t)^2 + Var(n_{ij}^t)\}} \tag{5}$$

Nous pouvons confirmer que $(\alpha^t)^*$ minimise les risques parce que $R^{"}(\alpha^t) \geq 0$ pour tous les α^t.

Le facteur d'oubli $(\alpha^t)^*$ conduit à la meilleure estimation en terme de minimisation des risques. Il nécessite la connaissance de la matrice de proximité réelle Ψ^t, qui est ce que nous essayons d'estimer, et la variance du bruit $Var(N^t)$.

Étant donné que nous traitons des adhésions des objets aux clusters, nous proposons de faire les hypothèses suivantes sur la structure de Ψ^t et $Var(N^t)$:

- $\Psi^t_{ii} = \Psi^t_{jj}$ pour deux objets quelconques i et j qui appartiennent au même cluster.
- $\Psi^t_{ij} = \Psi^t_{lm}$ pour deux objets distincts i, j et deux objets quelconques distincts l, m de telle sorte que i, l appartiennent au même cluster, et j, m appartiennent au même cluster.

2.2 Estimation adaptative du facteur d'oubli

Nous pouvons échantillonner sur chaque bloc afin d'estimer les entrées de Ψ^t et de $Var(N^t)$ et les remplacer pour obtenir une estimation $(\widehat{\alpha}^t)^*$ de $(\alpha^t)^*$. En effet, pour estimer les entrées de $\Psi^t = E[W^t]$, nous procédons comme suit :

- Pour deux objets distincts i et j du même cluster c, nous pouvons estimer W^t_{ij} en utilisant la moyenne d'échantillon donnée par l'équation (6).

$$\widehat{E}[W^t_{ij}] = \frac{1}{\mid c \mid (\mid c \mid -1)} \sum_{l \in c} \sum_{m \in c,\ m \neq l} W^t_{lm} \tag{6}$$

  De même, nous estimons Ψ^t_{ii} dans l'équation (7).

$$\widehat{E}[\Psi^t_{ii}] = \frac{1}{\mid c \mid} \sum_{l \in c} W^t_{ll} \tag{7}$$

- Pour les objets distincts i dans les clusters c et d avec $c \neq d$, nous estimons Ψ^t_{ij} dans l'équation (8).

$$\widehat{E}[W^t_{ij}] = \frac{1}{\mid c \mid\mid d \mid} \sum_{l \in c} \sum_{m \in d} W^t_{lm} \tag{8}$$

2.3 Fonction générique évolutive

La fonction générique évolutive est donnée comme suit :

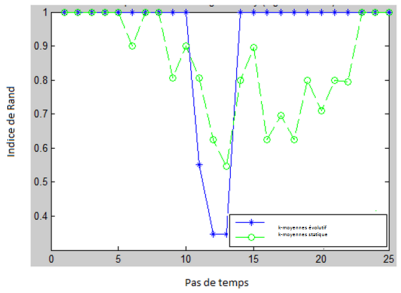
Algorithm 1 EvolClus

Entrée: Matrice de données
Sortie: Centres de clusters C^t au moment t

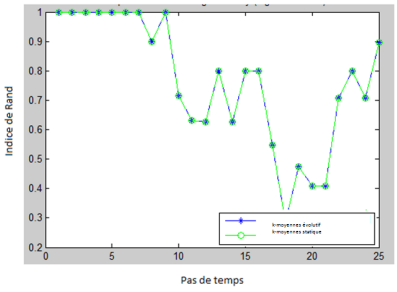
1: $C^t \Leftarrow C^{t-1}$
2: **Pour chaque** $i = 1, 2, ...$ **Faire**
3: Calculer $\widehat{E}[W^t]$
4: Calculer $\widehat{Var}(W^t)$
5: Calculer $(\widehat{\alpha}^t)^*$ par substitution des estimations $\widehat{E}[W^t]$ et $\widehat{Var}[W^t]$
6: $\widehat{W^t} \Leftarrow (\widehat{\alpha}^t)^* \widehat{W^{t-1}} + [1 - \widehat{\alpha}^t)^*] W^t$
7: **Fin Pour**
8: **Retourner** C^t

3 Expérimentation

Nous allons tester la capacité du cadre à s'adapter à un changement dans l'appartenance aux clusters. Pour ceci, nous allons varier les α^t et voir le comportement de deux cadres.

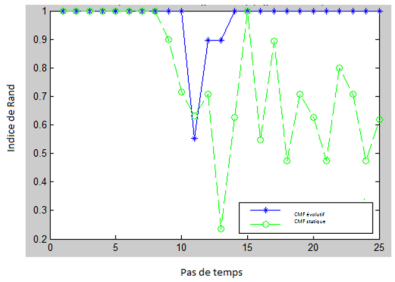


(**b**) $\boldsymbol{\alpha^t}$ estimé

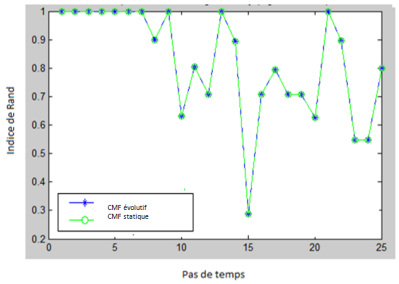


(**b**) $\boldsymbol{\alpha^t}$ fixe

FIG. 1 – *Variation de l'indice de Rand en fonction des pas de temps selon les cadres statique et évolutif des k-moyennes selon les cadres statique et évolutif avec α^t fixe et estimé.*

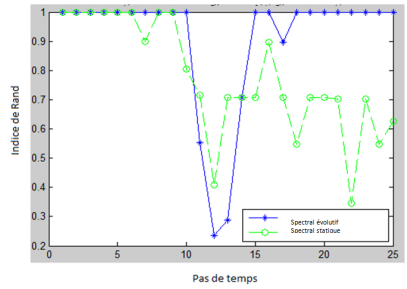


(**a**) $\boldsymbol{\alpha^t}$ estimé

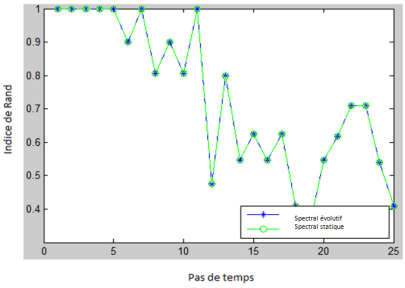


(**b**) α^t fixe

FIG. 2 – *Variation de l'indice de Rand en fonction des pas de temps selon les cadres statique et évolutif du CMF selon les cadres statique et évolutif avec α^t fixe et estimé.*



(**a**) $\boldsymbol{\alpha^t}$ estimé



(**b**) $\boldsymbol{\alpha^t}$ fixe

FIG. 3 – *Variation de l'indice de Rand en fonction des pas de temps selon les cadres statique et évolutif du clustering spectral avec α^t fixe et estimé.*

Pour les *k-moyennes*, la figure 2 montre que si nous estimons α^t, la valeur de l'indice de Rand (1971) est plus grande pour cadre évolutif. Par contre avec $\alpha^t = 0.5$, la figure 2 montre qu'il y a égalité entre l'algorithme des *k-moyennes* dans les cadres statique et évolutif.

D'après la figure 3, nous remarquons que lorsque nous estimons α^t, la valeur de l'indice de Rand est plus grande dans le cadre évolutif que celui statique. L'algorithme de *CMF* statique ne fonctionne pas bien dès que les clusters commencent à se chevaucher vers environ le $9^{\text{ème}}$ pas de temps. Avec $\alpha^t = 0.5$, les deux cadres fournissent les mêmes résultats.

Pour assurer le bon fonctionnement du cadre proposé, nous allons étendre sa comparaison avec les résultats de deux cadres *PCQ* et *PCM* de Chi et al. (2009). Le tableau 1 montre que le cadre proposé avec l'algorithme de *CMF* reste toujours plus efficace que les deux cadres *PCQ* et *PCM* puisque le résultat de la valeur de l'indice de Rand (1971) est toujours plus grand.

Méthode	Paramètre	Indice de Rand
CMF statique	-	0.796
CMF incrémental	estimation α^t	**0.963**
PCQ	a^t formés	0.910
	$a^t = 0.5$	0.823
PCM	a^t formés	0.842
	$a^t = 0.5$	0.810

TAB. 1 – *Etude comparative des indices de Rand entre PCM, PCQ et CMF incrémental.*

De même, nous comparons l'approche spectrale de clustering évolutif avec celle statique avec un α^t estimé et fixe. La figure 3 montre que l'algorithme fonctionne bien avec l'aspect incrémental et donne une valeur de l'indice de Rand (1971) plus grande que celle obtenue avec le cadre statique. Ce qui implique le bon fonctionnement du cadre avec les approches spectrales.

Avec $\alpha^t = 0.5$, nous remarquons, d'après 3, que le comportement est le même pour les deux cadres statique et évolutif.

4 Conclusion

Le cadre proposé dans ce travail surpasse généralement par celui statique en produisant des clusters qui reflètent les tendances à long terme tout en étant robuste aux variations à court terme. Il est universel dans le sens qu'il permet à n'importe quel algorithme de clustering statique d'être étendu à un caractère évolutif qui fournit une méthode explicite pour la sélection du facteur d'oubli, contrairement aux méthodes existantes. L'objectif était de suivre avec précision la matrice réelle de proximité à chaque pas de temps. Cela a été accompli en utilisant une mise à jour récursive avec un facteur d'adaptation d'oubli qui contrôle la quantité de poids à appliquer aux données historiques.

L'expérimentation a donné des résultats reflétant la performance du cadre adapté dans la performance de l'opération du clustering par rapport à celle statique.

Comme perspectives de travail, nous proposons d'élargir les expérimentations sur d'autres jeux de données et la possibilité d'étendre ce cadre afin qu'il puisse supporter des larges BD par réduction de l'échelle et l'échantillonnage des données.

Références

Bezdek, J. (1984). Fcm: The fuzzy c-means clustering algorithm. *Computers et Geosciences 10(2-3)*, 191–203.

Chakrabarti, D., R. Kumar, et A. Tomkins (2006). Evolutionary clustering. *In Proceedings of the* 12^{th} *ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, 554–560.

Chen, A., Y. Wiesel, C. Eldar, et A. Hero (2010). Shrinkage algorithms for mmse covariance estimation. *IEEE Transactions on Signal Processing 58(10)*, 5016–5029.

Chi, Y., X. Song, D. Zhou, K. Hino, et B. Tseng (2009). On evolutionary spectral clustering. *ACM Transactions on Knowledge Discovery from Data 3(4)*, 603–621.

Falkowski, T., J. Bartelheimer, et M. Spiliopoulou (2006). Mining and visualizing the evolution of subgroups in social networks. *In Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence*, 52–58.

Filippone, M., F. Camastra, F. Masulli, et S. Rovetta (2008). A survey of kernel and spectral methods for clustering. *Pattern recognition 41(1)*, 176–190.

Ledoit, O. et M. Wolf (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance 10(5)*, 603–621.

Mac-Queen, J. (1967). Some methods for classification and analysis of multivariate observations. *In Proceedings of the* 5^{th} *Berkeley Symposium on Mathematical Statistics and Probability*, 281–297.

Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association 66*, 846–850.

Schäfer, J. et K. Strimmer (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statist. Applic.Genetics Molec. Biology 4(32)*.

Tang, L., H. Liu, J. Zhang, et Z. Nazeri (2008). Community evolution in dynamicmulti-mode networks. *In Proceedings of the* 14^{th} *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 677–685.

Zhang, J., Y. Song, G. Chen, et C. Zhang (2009). On-line evolutionary exponential family mixture. *In Proceedings* 21^{st} *International Joint Conference on Artificial Intelligence*, 1610–1615.

Summary

Evolutionary clustering surpasses generally the static one by producing clusters that reflect the long-term trends while being robust to variations in the short term. Several algorithms have been proposed while adopting these aspects. They often operate by adding a fine temporal penalty cost function of a static clustering method. In this work, we adopt a different approach for evolutionary clustering by accurate monitoring of a proximity temporal variable between objects tracked by an ordinary static clustering.