Soft Subspace Growing Neural Gas pour le Clustering de Flux de Données

Mohammed Oualid Attaoui*,**, Mustapha Lebbah*, Nabil Keskes**
Hanene Azzag*, Mohammed Ghesmoune*

* University of Paris 13, Sorbonne Paris City LIPN-UMR 7030 - CNRS,
99 , av. J-B Clément, 93430 Villetaneuse, France
attaoui, mustapha.lebbah, azzag, ghesmoune@lipn.univ-paris13.fr
** Higher School of Computer Science (ESI-SBA), LabRI Laboratory, Sidi Bel-Abbes, Algeria
n.keskes@esi-sba.dz

Résumé. Le clustering de sous-espaces a été appliquée avec succès dans de nombreux domaines, son objectif est de détecter simultanément les clusters et les sous-espaces d'attributs d'origine dans lequel ces clusters existent. Un flux de données est une séquence massive de données venant en continu. Le clustering de ce type de données nécessite certaines restrictions de temps et de mémoire. Dans cet article, nous proposons une nouvelle méthode appelée S2G-Stream basée sur le clustering de flux de données et le clustering de sous-espaces souple. Des expériences sur des ensembles de données ont montré la capacité de S2G-Stream à détecter simultanément les meilleures attributs, sous-espaces et le meilleur clustering.

1 Introduction

Contrairement aux données traditionnelles qui sont immuables et statiques, un flux de données a ses propres caractéristiques : (1) il est infinie. (2) il est en évolution rapide et se produit en temps réel avec des exigences de réponse rapides. (3) un seul (ou peu) passage est possible à travers les données. (4) le stockage du flux de données est limité, seul un résumé des données peut être enregistré. Le clustering de flux de données est une technique qui effectue une analyse de cluster sur un flux de données et est capable de produire des résultats en temps réel. La capacité de traiter les données en un seul passage et de les résumer, tout en utilisant une mémoire limitée, est cruciale pour la clustering des flux de données. Subspace Clustering est une extension de la sélection d'attributs qui tente d'identifier les clusters dans différents sous-espaces du même jeu de données. En tant que sélection d'attributs, le subspace clustering nécessite à la fois une méthode de recherche et un critère d'évaluation. De plus, le clustering de sous-espaces doit restreindre d'une manière ou d'une autre la portée du critère d'évaluation afin de prendre en compte différents sous-espaces pour chaque cluster différent.

Dans le travail précédent Ouattara et al. (2013), les auteurs ont introduit 2S-SOM pour le subspace clustering basé sur les cartes auto-organisatrices SOM. Notre contribution principale dans cet article consiste à étendre le travail dans Ouattara et al. (2013) au clustering de flux de

données en se basant sur le modèle Growing Neural Gas. Nous avons introduit les notions de réservoir et de fading pour tenir compte de la nature évolutive des données de flux. Le reste de l'article est organisé comme suit : dans la section 2, nous présentons l'état de l'art. Dans la section 3, nous expliquons notre algorithme appelé Soft Subspace Clustering de flux de données (S2G-Stream) et évaluons ses performances dans la section 4. Enfin, nous concluons cet article dans la section 5.

2 Etat de l'art

2.1 Clustering de flux de données

Le clustering de flux de données est généralement effectué en tant que processus en deux étapes avec une partie en ligne qui résume les données dans de nombreux micro-clusters ou cellules de grille, puis dans un processus hors ligne, ces micro-clusters (cellules) sont regroupés/fusionnés dans un plus petit nombre de groupes finaux Ren et Ma (2009); Shukla et al. (2017). Plus récemment, les algorithmes utilisent également des stratégies d'apprentissage compétitives afin d'adapter les centroïdes des cluster au fil du temps. Ceci est inspiré par les cartes auto-organisatrices(SOM) Kohonen (2012) où des clusters entrent en concurrence pour représenter une observation. SOStream Isaksson et al. (2012) (Clustering basé sur la densité auto-organisée sur un flux de données) associe DBSCAN Ester et al. (1996) à des cartes auto-organisatrices (SOM) pour le clustering. Il introduit un seuil de similarité automatiquement adapté pour le clustering basée sur la densité afin de créer des voisinages avec un nombre minimal de points. Suivant l'idée d'un apprentissage compétitif, l'algorithme déplace également les k-voisins les plus proches du cluster. Si les clusters se rapprochent au cours de cette étape, ils seront fusionnées.

2.2 Subspace Clustering

En fonction de la manière dont les sous-espaces de cluster sont déterminés, le subspace clustering peut être classé en deux catégories principales : le hard subspace clustering (HSC) et le soft subspace clustering (SSC). Les algorithmes SSC effectuent un clustering dans des espaces de grande dimension en attribuant un poids à chaque dimension afin de mesurer la contribution de chaque dimension à la formation d'un cluster particulier Deng et al. (2016). Les méthodes de SSC peuvent être classées en trois catégories. CSSC (Subspace Clustering conventionnel) utilise un processus de pondération des attributs dans un clustering en deux étapes. Premièrement, il utilise des stratégies de pondération pour trouver des sous-espaces. Ensuite, le clustering est effectué sur le sous-espace obtenu. Ceci est appelé pondération séparée des attributs. Le clustering peut également être obtenue en effectuant les deux processus simultanément, on parle de pondération couplée des attributs. ISSC (Subspace Clustering Indépendant) associe à chaque cluster son propre vecteur de pondération afin que chaque cluster forme son propre sous-espace. XSSC (subspace clustering étendu) a été proposé pour améliorer les performances des CSSC et des ISSC.

Dans HSC, les attributs contribuent équitablement dans le processus de clustering. Les algorithmes HSC peuvent être divisés en méthodes de recherche ascendantes et déscendante Friedman et Meulman (2004). Les algorithmes descendants recherchent une classification ini-

tiale dans l'ensemble des dimensions et évaluent le sous-espace de chaque cluster. Les approches ascendantes définissent un histogramme pour chaque dimension, une plage avec une densité supérieure à un seuil fixe représente un cluster.

3 Modéle proposé

Dans cette section, nous présentons S2G-Stream une extension du modéle 2S-SOM Ouattara et al. (2013) pour le clustering de flux de données. Notre modéle se base sur le modèle de gas neuronal évolutif (GNG). Il s'agit d'une approche auto-organisatrice incrémentale appartenant à la famille des cartes topologiques telles que les cartes auto-organisatrices (SOM) Kohonen (2012). Nous supposons que le flux de données consiste en une séquence $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ de n (potentiellement infini) d'éléments arrivant au temps $t_1, t_2, ..., t_n$, où $\mathbf{x}_i = (x_i^1, x_i^2, ..., x_i^d)$. Dans chaque itération, S2G-Stream est représenté par un graphe $\mathcal{C}$, où chaque noeuds représente un cluster. Chaque noeud $c \in \mathcal{C}$ est associé à : (1) un prototype $\mathbf{w}_c = (w_c^1, w_c^2, \ldots, w_c^d)$ qui représente sa position (2) un poids π_c (3) une variable erreur $error(c)$ représentant la distance entre ce noeuds et les points associés à lui. Pour chaque paire de noeuds (r, c), nous désignons le chemin le plus court entre r et c sur le graphe par $\delta(c, r)$. Enfin on désigne par $\mathcal{K}^T(\delta) = \mathcal{K}(\delta/T)$ la fonction voisinage, T contrôle la largeur de $\mathcal{K}$.

En se basant sur les travaux Ouattara et al. (2013); Chen et al. (2012), nous introduisons un double système de pondération pour les attributs, noté β, et les sous-espaces, noté α, afin de faire en sorte que les attributs et les sous-espaces pertinents contribuent davantage au clustering. Les attributs $\mathcal{F}$ sont divisées en P sous-espaces, $\mathcal{F} = \cup_{b=1}^{P} \mathcal{F}_b$ où $\mathcal{F}_b = \{x^j, j = 1, ..., d_b\}$ où $d_1 + ... + d_b + ... + d_P = d$. Ainsi, α peut être représenté sous la forme d'une matrice $K \times P$ où α_c^b est le poids du sous-espace b dans le noeud c. β est une matrice $K \times d$ où β_b est une matrice $K \times d_b$, où $\beta_{cb}^j (j = 1, ..., d_b)$ est le poids de l'attribut j^{th} dans le sous-espace b pour le noeud c avec $\sum_{j=1}^{d_b} \beta_{cb}^j = 1$ et $\sum_{b=1}^{P} \alpha_c^b = 1, \forall c \in \mathcal{C}$. Les sous-espaces peuvent être extraits des deux matrices de pondération.

Nous proposons de minimiser la nouvelle fonction de coût définie ci-dessous pour chaque batch de données $\mathcal{X}^{(t+1)} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_{t+1}\}$:

$$\jmath^{(t+1)}(\phi, \mathcal{W}, \alpha, \beta) = \sum_{c \in C} \sum_{b=1}^{P} \sum_{\mathbf{x}_i \in \mathcal{X}^{(t+1)}} \mathcal{K}^T \left(\delta(c, \phi(\mathbf{x}_i))\right) \alpha_c^b \mathcal{D}_{\beta_{cb}} + J_{cb} + I_c \qquad (1)$$

Où $\mathcal{D}_{\beta_{cb}} = \sum_{j=1}^{d_b} \beta_{cb}^j (x_i^j - \omega_c^j)^2$. $I_c = \lambda \sum_{b=1}^{P} \alpha_c^b \log(\alpha_c^b)$ and $J_{cb} = \eta \sum_{j=1}^{d_b} \beta_{cb}^j \log(\beta_{cb}^j)$ représentent respectivement les entropies négatives pondérées associées aux vecteurs de pondération des sous-espaces et des vecteurs de pondération des attributs. Les paramètres λ et η permettent d'ajuster les contributions relatives apportées par les attributs et les sous-espaces au clustering.

3.1 Algorithme d'optimisation

L'optimisation de la fonction de coût s'effectue pour chaque batch de données $\mathcal{X}^{(t+1)}$ en quatre étapes correspondant aux quatre paramètres $\mathcal{W}, \phi, \alpha$ and β :

1. **Fonction d'affectation :** Pour $\mathcal{W}, \alpha$ et β fixes, la fonction d'affectation $\phi(\mathbf{x}_i)$ est décrite dans l'équation (2). Afin de réduire le temps de calcul, les nœuds de voisinage ne sont pas pris en compte dans l'affectation.

$$\phi(\mathbf{x}_i) = \arg\min_{c \in \mathcal{C}} \left(\sum_{b=1}^{P} \alpha_c^b \sum_{j=1}^{d_b} \beta_{cb}^j \left(x_i^j - \omega_c^j \right)^2 \right) \tag{2}$$

2. **Mise à jours des prototypes $\mathcal{W}$:** Pour $\mathcal{W}, \alpha$ et β fixes, $\mathbf{w}_c$ prototypes sont mis à jour pour chaque batch de données suivant l'équation définie ci-dessous.

$$\mathbf{w}_c^{(t+1)} = \frac{\mathbf{w}_c^{(t)} n_c^{(t)} \gamma + \sum_{r \in C} \mathcal{K}^T(\delta(r,c)) \mathbf{w}_r^{(t)} m_r^{(t)}}{n_c^{(t)} \gamma + \sum_{r \in C} \mathcal{K}^T(\delta(r,c)) m_r^{(t)}} \tag{3}$$

où $\mathbf{w}_c^{(t)}$ est le prototype précédent, $n_c^{(t)}$ est le nombre de points attribués au cluster, $\mathbf{w}_r^{(t)}$ est le prototype précédent du cluster r (qui est un voisin de c) et $m_r^{(t)}$ est le nombre de points ajoutés au cluster r dans le lot actuel : $n_c^{(t+1)} = n_c^{(t)} + m_c^{(t)}$.

3. **Mise à jours des poids α :** pour ϕ, $\mathcal{W}$ et β fixes, nous mettons à jours les poids α comme suit :

$$\alpha_c^b = \frac{e^{\frac{-D_{cb}}{\lambda}}}{\sum_{s=1}^{P} e^{\frac{-D_{cs}}{\lambda}}} \tag{4}$$

with $D_{cb} = \sum_{\mathbf{x}_i \in \mathcal{X}^{(t)}} \mathcal{K}^T(\delta(\phi(\mathbf{x}_i), c)) \sum_{j=1}^{d_b} \beta_{cb}^j \left(\mathbf{x}_i^j - w_c^j \right)^2$

4. **Mise à jours des poids β :** pour ϕ, $\mathcal{W}$ et α fixes, nous mettons à jours les poids β comme suit :

$$\beta_{cb}^j = \frac{e^{\frac{-E_{cb}^j}{\eta}}}{\sum_{h \in \mathcal{F}_{P_j}} e^{\frac{-E_{ch}}{\eta}}} \tag{5}$$

avec $E_{cb}^j = \sum_{\mathbf{x}_i \in \mathcal{X}^{(t)}} \alpha_c^{P_j} \mathcal{K}^T(\delta(\phi(\mathbf{x}_i), c)) \left(\mathbf{x}_i^j - w_c^j \right)^2$, où P_j est le sous-espace du j-ème attribut.

3.2 S2G-Stream algorithm

S2G-Stream est une extension de 2S-SOM pour les flux de données en se basant sur la topologie évolutive du modéle GNG. En commençant par deux noeuds et chaque fois qu'un nouveau point de données est disponible, nous relions le nœud le plus proche et le deuxième plus proche par une arête. Le nœud le plus proche et ses voisins topologiques sont déplacés vers le point de donnée. La description complète de l'algorithme S2G-Stream se trouve dans Algorithme 1.

Algorithme 1 : S2G-Stream

input : $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}, \pi_{min}, \tau_{age}, age_{max}, d, \eta, \lambda, \mu, \gamma, b$
output : prototypes :$W = w_1, w_2, .., w_n$ matrices de poids d'attributs β et de sous-espaces α
Initialiser le graphe avec deux nœuds, initialiser les poids α et β aléatoirement ;
while *il y a un batch de données* **do**
- Obtenir le batch de points de données arrivé à l'instant t ;

1. **Etape d'affectation**
   - Rechercher les nœuds bmu_1 et bmu_2 le plus proche et le deuxième plus proche en utilisant l'équation (2) et créer une arête entre eux. Si elle existe déjà : mettre l'âge à zéro ;
   -Affecter chaque point au centre le plus proche en utilisant l'équation (2) ;
2. **Etape de mise à jours** :
   Mettre à jours les prototypes comme décrit dans (3) ;
3. **Gestion des arêtes** : Algorithme (2)
4. **Mise à jours des poids** : Mettre à jour les poids des attributs (4) et les poids des sous-espaces (5);

Algorithme 2 : Gestion des arêtes

- Incrémenter l'âge de toutes les arêtes émanant du prototype le plus proche et les pondère ;
- Supprimer les arêtes dont l'âge est supérieur à age_{max} et mettre à jours $error$ de chaque noeuds : $error(bmu_1) = error(bmu_1) + ||\mathbf{x}_i - bmu_1||^2$;
- Appliquer la fonction fading comme $\pi_c^{(t+1)} = \pi_c^{(t)}\gamma$ et supprimer les noeuds dépassés
- Trouver le noeud q avec l'erreur la plus grande et son voisin f avec l'erreur accumulée la plus grandeet ajouter un noeud r entre eux : $\mathbf{w}_r = 0.5(\mathbf{w}_q + \mathbf{w}_f)$;
- Réduire les variables d'erreur de q et f en les multipliant par une constante υ où $0 < \upsilon < 1$ et affecter à r la valeur d'erreur de q ;
- Diminuer l'erreur de tous les nœuds en les multipliant par une constante s, et supprimer les nœuds isolés ;

4 Résultats expérimentaux

4.1 Evaluation du clustering

Nous avons évalué la qualité du clustering de S2G-Stream sur des jeux de données réels de l'UCI Frank et Asuncion (2010) décrits dans le tableau 1. Pour les mesures de qualité, nous avons utilisé l'information normalisée mutuelle (NMI) et l'indice de RAND ajusté (ARAND).

En supposant que des données de grande dimension arrivent en continu, S2G-Stream divise les données en continu en batchs et traite chaque batch en continu. La taille du batch dépend de la mémoire disponible et de la taille du jeu de données d'origine. Nous définissons l'intervalle de temps entre deux batch à 1 seconde. Nous avons répété nos expériences avec différentes initialisations et avons choisi celles qui donnent les meilleurs résultats. Nous avons défini $\mu = 3$, $\gamma = 0.99$ et $age_{max} = 250$.

Pour montrer l'efficacité de notre méthode, nous la comparons à quatre algorithmes : Growing Neural Gas (GNG) de la librairie Smile [1], $CluStream$ et $DStream$ de la librairie StreamMOA [2]. Les résultats sont présentés dans le tableau 1. Il est à noter que S2G-Stream donne de meilleurs résultats que les autres méthodes, à l'exception de DStream sur CTG et $waveform$ avec la métrique NMI et de CTG avec la métrique ARAND. Ces résultats sont dus au fait que S2G-Stream détecte les attributs et les sous-espaces pertinents et permet à ces sous-espaces de contribuer davantage au clustering. Cela est également dû à la notion de fading qui réduit l'impact de données non significatives.

Dataset	Mesures	S2G-Stream	GNG	CluStream	DStream
waveform	NMI	0.397±0.002	0.306±0.078	0.393±0.065	**0.434±0.003**
	ARAND	**0.137±0.007**	0.006±0.103	0.010±0.001	0.040±0.001
IS	NMI	**0.550±0.05**	0.542±0.010	0.506±0.065	0.435±0.07
	ARAND	**0.418±0.04**	0.102±0.051	0.098±0.010	0.134±0.002
CTG	NMI	0.270±0.009	0.375±0.004	0.086±0.06	**0.471±0.170**
	ARAND	0.124±0.005	0.030±0.011	0.019±0.008	**0.209±0.002**
pendigits	NMI	**0.672±0,038**	0.585±0.019	0.285±0.099	0.554±0.15
	ARAND	**0.408±0.060**	0.027±0.085	0.011±0.006	0.016±0.011

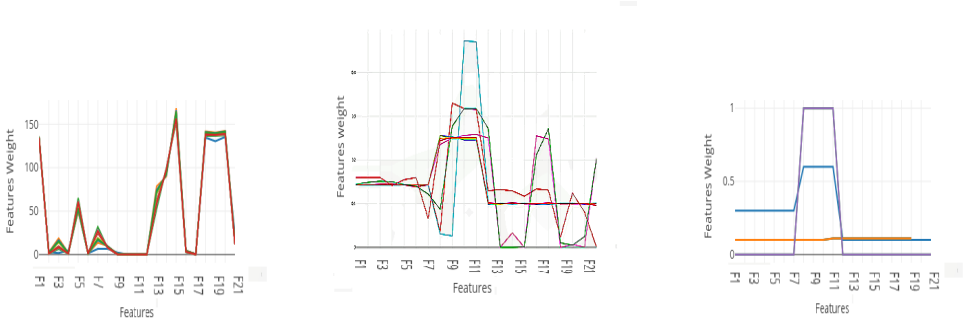
TAB. 1: Comparaison de S2G-Stream avec différents algorithmes. La valeur après ± correspond à l'écart-type.

4.2 Analyse des sous-espaces et des attributs

Pour cette section, nous nous concentrons sur le jeu de données CTG. Le jeu de données *CTG* décrit les cardiotocogrammes de fœtus et est composé de 3 sous-espaces, le sous-espace 1 contient 7 variables liées à la fréquence cardiaque du fœtus. Le sous-espace 2 contient quatre variables décrivant la variabilité de la fréquence cardiaque. Le sous-espace 3 est composé de 10 variables définissant les histogrammes de la cardiographie fœtale.

1. http ://haifengl.github.io/smile/
2. https ://github.com/mhahsler/streamMOA

La figure 1 représente les prototypes $\mathcal{W}$, poids β et α pour le dernier batch de données CTG. Nous observons dans la figure (1b) les poids des attributs (8,9,10,11) qui sont respectivement ASTV (pourcentage de temps avec variabilité anormale à court terme), MSTV (valeur moyenne de la variabilité à court terme), ALTV (pourcentage du temps avec une variabilité à long terme anormale) et MLTV (valeur moyenne de la variabilité à long terme) sont plus élevés que les poids des autres attributs pour la plupart des clusters. Nous observons que ces 4 attributs influencent significativement le processus de clustering et sont plus importants que les autres attributs pour la plupart des clusters. Dans la figure (1c), nous observons que la pondération α du deuxième sous-espace contenant ces quatre caractéristiques est également supérieure à celle des deux autres sous-espaces. Nous concluons de cette expérience que la variabilité de la fréquence cardiaque influence mieux la classification des cardiotocogrammes fœtaux.



(a) Prototypes $\mathcal{W}$ (b) β weights (c) α weights

FIG. 1: Résultats des poids α et β, et prototypes $\mathcal{W}$ pour le dernier batch du jeux de données CTG. Chasue couleur représente un noeud.

5 Conclusion

Dans cet article, nous avons proposé S2G-Stream, une méthode efficace pour le subspace clustering de flux de données. Nous avons étendu les travaux précédents vers les données infinies en forme de flux en se basant sur la topologie évolutive de GNG. Nous avons également introduit la notion de fading pour supprimer les nœuds obsolètes. L'évaluation expérimentale et la comparaison avec des méthodes de classification bien connues démontrent l'efficacité et l'efficience de S2G-Stream en ce qui concerne les résultats du clustering, la découverte de clusters de forme arbitraire et la détection des attributs et des sous-espace. Les perspectifs incluent plusieurs expériences en modifiant l'ordre des données et en généralisant notre méthode pour traiter différents types de données. Nous prévoyons également de visualiser nos résultats en fonction de la topologie de GNG.

Références

Chen, X., Y. Ye, X. Xu, et J. Z. Huang (2012). A feature group weighting method for subspace clustering of high-dimensional data. *Pattern Recognition 45*(1), 434–446.
en

Deng, Z., K.-S. Choi, Y. Jiang, J. Wang, et S. Wang (2016). A survey on soft subspace clustering. *Information Sciences 348*, 84–106.

Ester, M., H.-P. Kriegel, J. Sander, X. Xu, et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, Volume 96, pp. 226–231.

Frank, A. et A. Asuncion (2010). Uci machine learning repository [http ://archive. ics. uci. edu/ml]. irvine, ca : University of california. *School of information and computer science 213*.

Friedman, J. H. et J. J. Meulman (2004). Clustering objects on subsets of attributes (with discussion). *Journal of the Royal Statistical Society : Series B (Statistical Methodology) 66*(4), 815–849.

Isaksson, C., M. H. Dunham, et M. Hahsler (2012). Sostream : Self organizing density-based clustering over data stream. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pp. 264–278. Springer.

Kohonen, T. (2012). *Self-organizing maps*, Volume 30. Springer Science & Business Media.

Ouattara, M., N. N. Keita, F. Badran, et C. Mandin (2013). Soft subpace clustering pour données multiblocs basée sur les cartes topologiques auto-organisées som : 2s-som. In *SFDS 2013*.

Ren, J. et R. Ma (2009). Density-based data streams clustering over sliding windows. In *Fuzzy Systems and Knowledge Discovery, 2009. FSKD'09. Sixth International Conference on*, Volume 5, pp. 248–252. IEEE.

Shukla, M., Y. Kosta, et M. Jayswal (2017). A modified approach of optics algorithm for data streams. *Engineering, Technology & Applied Science Research 7*(2), 1478–1481.

Summary

Subspace clustering has been successfully applied in many domains and its goal is to simultaneously detect both clusters and subspaces of the original feature space where these clusters exist. A Data stream is a massive sequences of data coming continuously. Clustering this type of data requires some restrictions in time and memory. In this paper we propose a new method named S2G-Stream based on clustering data streams and soft subspace clustering. Experiments on public datasets showed the ability of S2G-Stream to detect simultaneously the best features, subspaces and the best clustering.