Subspace Clustering et Visualisation des Flux de Données

Ibrahim Louhi[*,**] Lydia Boudjeloud-Assala[*]
Thomas Tamisier[**]

[*]Université de Lorraine, Laboratoire d'Informatique Théorique et Appliquée.
{ibrahim.louhi, lydia.boudjeloud-assala}@univ-lorraine.fr
[**]Luxembourg Institute of Science and Technology.
{ibrahim.louhi, thomas.tamisier}@list.lu

Résumé. Dans ce papier nous proposons une nouvelle approche de subspace clustering pour les flux de données, permettant à l'utilisateur de suivre visuellement le changement dans le comportement du flux. Cette approche détecte l'impact des variables sur l'évolution du flux, Tout en visualisant les étapes du subspace clustering en temps réel. En premier lieu nous appliquons un clustering sur l'ensemble de variables afin d'identifer les sous-espaces. Ensuite un clustering est appliqué sur les individus dans chaque sous-espace.

1 Introduction

Le clustering est une des techniques utilisées pour la fouille de données, qui essaye de regrouper les individus similaires selon certains critères dans le même groupe appelé cluster. Cependant, parfois les données comportent des informations cachées qui ne sont pas visibles sur l'espace original de variables. Parmi les techniques utilisées pour découvrir ces informations, le subspace clustering cherche à trouver des clusters sur tous les sous-espaces de données.

La tâche du subspace clustering se complique encore plus quand il s'agit de flux de données. Comment identifier des clusters dans ses sous-espaces pertinents tout en respectant les contraintes du traitement du flux de données. Plusieurs approches ont été proposées pour effectuer du subspace clustering sur les flux de données. Cependant aucune à notre connaissance ne permet de visualiser en temps réel l'évolution du flux et des sous-espaces. La visualisation permet dans le contexte du subspace clustering de mieux comprendre les résultats obtenus, et le plus important, d'explorer les données au niveau des différents sous-espaces.

Dans ce papier nous présentons dans un premier temps un bref état de l'art de quelques techniques utilisées pour le subspace clustering sur des données statiques, la visualisation des sous-espaces, et le subspace clustering appliqué aux flux de données. Ensuite, nous présentons notre approche pour effectuer et visualiser un subspace clustering de flux. Nous discuterons également les résultats obtenus, et nous illustrons l'utilité de notre approche et des perspectives pouvant l'améliorer.

2 Etat de l'art

Le subspace clustering essaye de trouver tous les clusters possibles sur tous les sous-espaces, tout en identifiant le meilleur sous-espace pour chaque cluster. CLIQUE (Agrawal et al., 1999) et ses extensions ENCLUS (Cheng et al., 1999) et MAFIA (Goil et al., 1999), combinent un clustering basé sur la densité et un clustering basé sur les grilles. Ils définissent d'abord les sous-espaces, pour ensuite chercher les unités denses adjacentes dans les grilles de chaque sous-espace. Les clusters sont formés par la combinaison de ces unités.

Dans le contexte des flux de données, une adaptation des techniques classiques de subspace clustering est nécessaire. DUCStream (Gao et al., 2005) se base sur l'algorithme CLIQUE (Agrawal et al., 1999). Après l'identification des clusters, DUCSTREAM effectue ensuite une mise à jour incrémentale des unités. HPSTREAM (Aggarwal et al., 2004) est une adaptation de Clustream (Aggarwal et al., 2003) qui est un algorithme de clustering pour les flux de données. HPSTREAM utilise un micro-clustering pour stocker un résumé statistique sur le flux de données (les clusters et leur position temporelle dans le flux), et un macro-clustering qui utilise ce résumé de données pour fournir le résultat obtenu par le clustering à n'importe quel point temporel du flux. Les clusters sont obtenus sur des sous-espaces, et chaque sous-espace est continuellement réévalué ce qui peut changer la structure des clusters déjà obtenus. Contrairement à HPStream qui fournit un résultat approximatif en se basant sur un résumé, INCPREDECON (Kriegel et al., 2011) a besoin d'accéder aux données brutes (un accès limité à un sous-ensemble des données seulement). Il fournit une meilleure solution équivalente au résultat obtenu d'une manière statique (traiter tout l'ensemble de données). Le principe de INCPREDECON consiste à mettre à jour les précédents clusters et leurs variables en se basant sur les nouvelles données.

Les dernières années plusieurs approches visuelles ont été proposées pour le subspace clustering telles que VISA (Assent et al., 2007), HEIDI MATRIX (Vadapalli et Karlapalem, 2009) et SUBVIS (Hund et al., 2016). Cependant, à notre connaissance il n'existe pas d'outils pour trouver et visualiser les sous-espaces dans le contexte des flux de données. Dans ce papier nous proposons une technique pour trouver automatiquement les sous-espaces dans un flux de données, et de visualiser les résultats obtenus dans le but de trouver des informations intéressantes qui n'étaient pas visibles sur la totalité de l'espace des variables.

3 Le subspace Clustering

Cette approche est une extension de l'algorithme NNG-Stream (Louhi et al., 2016) pour le traitement des flux de données (NNG : Nearest Neighborhood Graph). Plutôt que de traiter chaque nouvel individu individuellement dès son arrivée, NNG-Stream traite chaque groupe de nouveaux individus G_i simultanément. La taille des groupes $|G_i| = n$ est fixée par l'utilisateur suivant son expertise et ses préférences. Les clusters obtenus sur chaque nouveau groupe sont reliés avec les clusters globaux du flux selon une mesure de distance entre les centres de gravité des clusters. Chaque cluster est représenté par un graphe de voisinage.

Dans ce qui suit nous adaptons NNG-Stream pour le subspace clustering du flux en lui permettant de chercher des clusters définis sur des sous-espaces (sous-ensemble de variables), et de prendre en considération l'évolution du flux.

Soit $E = \{e_1, e_2, ...\}$ l'ensemble des individus du flux F et $D = \{d_1, ..., d_m\}$ est l'ensemble des variables des individus. Dès l'arrivée du premier groupe d'individus G_1 nous appliquons un algorithme de clustering basé sur le voisinage sur l'ensemble des variables D. Nous calculons la distance entre chaque couple de variables, deux variables sont considérées comme étant voisines seulement si leur distance est inférieure à un seuil. Chaque groupe de voisins représente un cluster, et chaque cluster représente un sous-espace de données. Ensuite pour chaque sous-espace obtenu, nous appliquons le clustering basé sur le voisinage sur les individus définis uniquement avec les variables du sous-espace.

A l'arrivée du groupe suivant G_2 , nous appliquons également le clustering basé sur le voisinage sur l'ensemble des variables D. Deux possibilités peuvent se présenter, soit nous obtenons les mêmes sous-espaces (les mêmes clusters de variables) que sur le premier groupe, soit les sous-espaces sont différents.

Si les sous-espaces restent inchangés, nous traitons les individus de ce deuxième groupe sur chaque sous-espace de la même manière que sur le groupe précédent G_1 et indépendamment des résultats obtenus sur ce dernier. Ensuite les nouveaux clusters sont utilisés pour mettre à jour les précédents clusters. Pour chaque sous-espace nous calculons la distance entre les médoides des nouveaux et des précédents clusters. Si deux médoides sont proches selon une mesure de distance, leurs clusters respectifs sont reliés. Dans le cas ou un nouveau cluster n'est proche d'aucun des clusters précédents, il est rajouté comme étant un nouveau cluster dans le flux. Ainsi de suite tant que les sous-espaces ne changent pas, nous continuons à traiter le flux groupe par groupe et à mettre à jour les précédents clusters.

Si à l'arrivée d'un nouveau groupe, nous obtenons des sous-espaces différents, nous considérons que le flux de données a changé. Nous ne pouvons plus mettre à jour les précédents clusters vu que les sous-espaces ont changé. Cela représente la fin de la première fenêtre, nous sauvegardons un résumé de la première fenêtre contenant le nombre et le contenu des sous-espaces ainsi que le nombre de clusters obtenus dans chaque sous-espace. Une fenêtre représente une partie du flux $T_i \rightarrow T_j$ avec les mêmes sous-espaces.

Nous traitons les individus des groupes de la deuxième fenêtre exactement de la même façon que la première. A chaque fois que les sous-espaces changent par rapport au groupe précédent ça représente la fin de la fenêtre actuelle, et à la fin de chaque fenêtre nous sauvegardons un résumé de ses sous-espaces et clusters. Le résumé permet de garder une trace sur le changement du flux.

4 Visualisation, résultats et discussion

Notre interface visuelle comporte plusieurs niveaux, une visualisation globale du flux (figure 1), une visualisation des sous-espaces (figure 2), une visualisation globale des clusters obtenus sur chaque sous-espace (figures 3 à 6) et une visualisation détaillée des clusters sur chaque sous-espace sous la forme de graphes de voisinage (Louhi et al., 2016).

Dans la figure 1 une partie du flux de données est représentée par une visualisation inspirée par les themerivers. Les lignes du themeriver représentent le nombre des clusters, le nombre d'outliers et le pourcentage d'outliers par rapport au nombre d'individus (le pourcentage est normalisé par rapport au nombre de clusters et d'outliers) à chaque instant T_i (les instants T_i représentent la fin de traitement de chaque groupe d'individus G_i). Nous avons choisi de re-

présenter seulement ces informations afin d'avoir une visualisation simple sans trop de détails, permettant à l'utilisateur de suivre l'évolution du flux sans un grand effort cognitif.

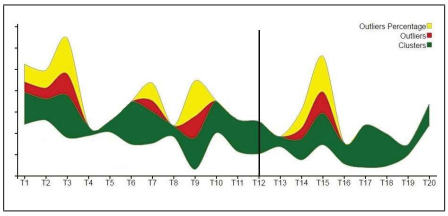

FIG. 1: Vue globale du flux de données

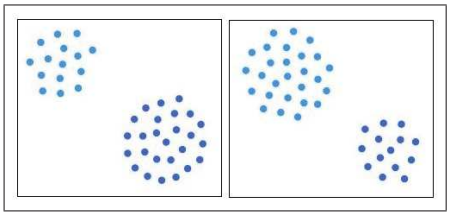

FIG. 2: Les deux sous-espaces obtenus

A la fin de chaque fenêtre (quand les sous-espaces obtenus changent), une ligne verticale s'affiche sur le themeriver (dans notre exemple ça se produit à l'instant T_{12}). Notre approche de subspace clustering applique un clustering sur l'ensemble des variables afin d'identifier les sous-espaces. Ces derniers peuvent être visualisés en même temps que la vue globale du flux (figure 2), où Chaque point représente une variable et chaque cluster représente un sous-espace. Ensuite un clustering est appliqué sur les individus de chaque sous-espaces. Une description du flux par un themeriver peut être obtenue (figures 3 à 6).

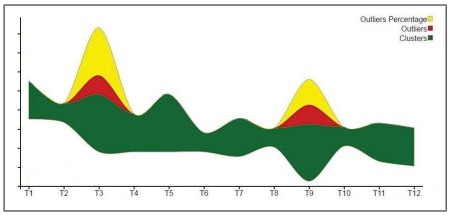

FIG. 3: La première fenêtre du flux sur le premier sous-espace

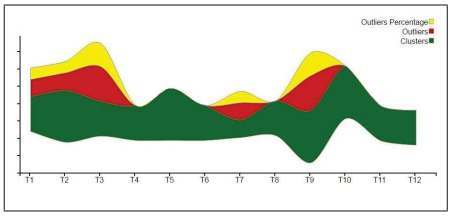

FIG. 4: La première fenêtre du flux sur le deuxième sous-espace

Les figures 3 et 4 représentent une description du clustering de la première fenêtre (T_1 jusqu'à T_{12}) sur les deux sous-espaces séparément. En comparant ces résultats avec le clustering sur tout l'espace, nous remarquons que sur le premier sous-espace (figure 3) il y a des outliers aux mêmes instants que sur le clustering global (T_3 et T_9) et que des outliers ont disparu à deux autres instants (T_1 et T_6). Sur le deuxième sous-espace (figure 4), nous retrouvons des outliers aux mêmes instants que sur l'espace original (T_1, T_3, T_6 et T_9).

A partir des themerivers, l'utilisateur peut afficher en détail les clusters obtenus à un instant T_i. Les clusters sont représentés sous la forme de graphes de voisinage afin de refléter l'algorithme de traitement. Cette visualisation des clusters a permis de comparer les outliers obtenus sur l'espace original et ceux obtenus sur les sous-espaces : aux instants T_3 et T_9 deux outliers sont détectés sur l'espace original à chacun des deux instants, seulement un des deux outliers est détecté sur le premier sous-espace à chaque instant. Sur le deuxième sous-espace, deux outliers sont détectés à chaque instant et ils sont identiques à ceux sur l'espace original. Aux instants T_1 et T_6 se sont exactement les mêmes outliers sur le deuxième sous-espace que sur l'espace original (un outlier à chaque instant).

Nous remarquons également que le deuxième sous-espace est assez proche de l'espace original, Le themeriver du deuxième sous-espace ressemble fortement à celui de la première partie du flux sur l'espace original des variables.

Les figures 5 et 6 représentent le flux de données sur la deuxième fenêtre (après T_{12}) sur les deux sous-espaces. En comparant ces résultats avec le clustering sur tout l'espace, nous remarquons que sur le premier sous-espace (figure 5) il y a des outliers au même instant que sur le clustering global (à l'instant T_{15}) et un nouveau outliers qui a apparu à T_{18}. Sur le deuxième sous-espace (figure 6), Les outliers ont disparu à T_{15} et des nouveaux sont apparus à T_{14}. La visualisation des clusters sous la forme de graphes de voisinage a permis également de comparer les outliers détectés. A l'instant T_{14} le même outlier est détecté sur l'espace original et sur le deuxième sous-espace. A l'instant T_{15} seulement un des deux outliers détectés sur l'espace original est détecté sur le premier sous-espace.

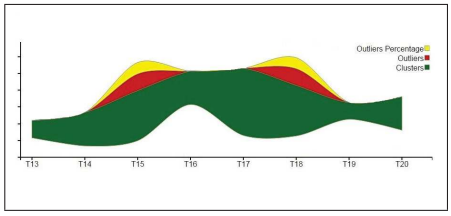

FIG. 5: La deuxième fenêtre du flux sur le premier sous-espace

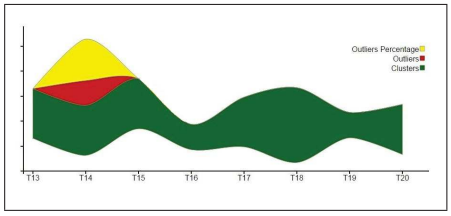

FIG. 6: La deuxième fenêtre du flux sur le deuxième sous-espace

A partir de ces visualisations (figures de 1 à 6) nous pouvons clairement comprendre l'intérêt de notre approche de subspace clustering pour les flux de données. Appliquer un clustering sur les variables permet de regrouper les variables qui ont le même degré d'influence sur le flux dans le même cluster. Ceci était visible quand nous avons pu détecter les mêmes outliers sur les sous-espaces que ceux sur l'espace original. Nous avons aussi pu trouver un sous-espace sur lequel le flux a le même comportement que sur l'espace original (figure 4). Nous pouvons imaginer l'intérêt de représenter un flux par un sous-espace dans les données à grande dimensionnalité, permettant d'optimiser le processus de traitement on ignorant les variables non pertinentes. Nous avons également détecté de nouveaux outliers sur des sous-espaces alors qu'ils n'apparaissaient pas sur l'espace original. Ce qui veut dire que nous avons découvert une information qui n'était pas visible dans l'espace original.

5 Conclusion

Dans ce papier nous avons proposé une nouvelle approche de subspace clustering pour les flux de données. La visualisation de toutes les étapes du subspace clustering a permis de mettre en évidence l'efficacité de cette approche. Nous avons pu trouver des sous-espaces qui représentent l'espace original de variables, un sous-espace sur lequel le flux à un comportement différent (de nouvelles informations sont visibles), et le plus important, nous avons pu détecter le changement dans le flux sous un nouvel angle. Plutôt d'identifer le changement par des tests statistiques, cela est possible en se focalisant sur l'évolution de l'impact des variables.

Références

Aggarwal, C. C., J. Han, J. Wang, et P. S. Yu (2003). A framework for clustering evolving data streams. In *Proceedings of the 29th international conference on Very large data bases-Volume 29*, pp. 81–92. VLDB Endowment.

Aggarwal, C. C., J. Han, J. Wang, et P. S. Yu (2004). A framework for projected clustering of high dimensional data streams. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pp. 852–863. VLDB Endowment.

Agrawal, R., J. E. Gehrke, D. Gunopulos, et P. Raghavan (1999). Automatic subspace clustering of high dimensional data for data mining applications. US Patent 6,003,029.

Assent, I., R. Krieger, E. Müller, et T. Seidl (2007). Visa : visual subspace clustering analysis. *ACM SIGKDD Explorations Newsletter 9*(2), 5–12.

Cheng, C.-H., A. W. Fu, et Y. Zhang (1999). Entropy-based subspace clustering for mining numerical data. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 84–93. ACM.

Gao, J., J. Li, Z. Zhang, et P.-N. Tan (2005). An incremental data stream clustering algorithm based on dense units detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 420–425. Springer.

Goil, S., H. Nagesh, et A. Choudhary (1999). Mafia : Efficient and scalable subspace clustering for very large data sets. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 443–452. ACM.

Hund, M., D. Böhm, W. Sturm, M. Sedlmair, T. Schreck, T. Ullrich, D. A. Keim, L. Majnaric, et A. Holzinger (2016). Visual analytics for concept exploration in subspaces of patient groups. *Brain Informatics*, 1–15.

Kriegel, H.-P., P. Kröger, I. Ntoutsi, et A. Zimek (2011). Density based subspace clustering over dynamic data. In *International Conference on Scientific and Statistical Database Management*, pp. 387–404. Springer.

Louhi, I., L. Boudjeloud-Assala, et T. Tamisier (2016). Traitement de flux par un graphe de voisinage incrémental. *Revue des Nouvelles Technologies de l'Information Fouille de Données Complexes, RNTI-E-31*, 15–36.

Vadapalli, S. et K. Karlapalem (2009). Heidi matrix : nearest neighbor driven high dimensional data visualization. In *Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery*, pp. 83–92. ACM.

Summary

In this paper we propose a novel subspace clustering approach for data streams, allowing the user a visual tracking of the data stream behavior. The approach detects the variables impact on the stream evolution. The subspace clustering steps are visualized on real time. First we apply a clustering on the variables set to obtain subspaces. Then we cluster the elements within each subspace.