Clustering visuel semi-interactif

Lydia Boudjeloud-Assala*, Philippe Pinheiro**,
Alexandre Blansché*, Thomas Tamisier** et Benoît Otjacques**

* Laboratoire d'Informatique Théorique et Appliquée, LITA-EA 3097,
Université de Lorraine, Metz, F-57045, France
prenom.nom@univ-Lorraine.fr
**LIST- Luxembourg Institute of Science and Technology,
Esch-Sur-Alzette, Luxembourg
prenom.nom@list.lu

Résumé. Nous proposons dans cet article une approche de clustering visuel semi-interactif. L'approche proposée utilise la perception visuelle pour guider l'utilisateur dans le processus interactif. Les clusters sont extraits de manière successive et itérative, puis évalués selon leur ordre d'extraction. Pour l'utilisateur, l'approche semi-interactive permet non seulement d'évaluer les classes en fonction d'un critère déterminé mais aussi d'évaluer l'influence de l'extraction d'un cluster sur ceux précédemment extraits. Un protocole de test est présenté afin de comparer cette approche avec les approches purement automatiques et purement interactives. Cet article est un résumé d'un papier accepté [1] pour un journal international.

1 Introduction

Dans le processus d'extraction de connaissances à partir de données, il y a au moins deux moyens de faire collaborer les méthodes automatiques avec des méthodes visuelles interactives. Il est possible d'utiliser les méthodes de visualisation en prétraitement de l'algorithme automatique ou en post-traitement de ce même algorithme. En prétraitement de données, on s'aperçoit que, bien souvent, une intuition initiale des concepts cachés peut être acquise de façon visuelle dans les très grandes quantités d'information. Cette étape peut également guider l'utilisateur dans le choix des algorithmes de fouille les plus pertinents ou de leurs paramètres. En post-traitement des connaissances, les méthodes de visualisation sont plutôt utilisées pour interpréter et évaluer des résultats en se basant sur des représentations graphiques plus accessibles que des colonnes de chiffres ou un ensemble de règles. Ces différentes interactions illustrent l'intérêt de faire coopérer des méthodes automatiques et des méthodes visuelles interactives. La compréhension des résultats est ainsi accrue et la précision des algorithmes automatiques peut être facilement améliorée. Une des possibilités pour augmenter la part de la visualisation dans les algorithmes de fouille de données est de faire coopérer l'algorithme automatique de fouille de données avec un algorithme visuel interactif. On parle alors de fouille

1. Interactive and iterative visual clustering, Information Visualization Journal, pp 1-17, 2015 (DOI: 10.1177/1473871615571951)

visuelle de données qui se distingue de la visualisation d'information et consiste donc, en l'utilisation de la visualisation comme outil pour assister la fouille. L'approche proposée dans cet article s'inscrit dans cette direction, nous nous basons sur les projections des données en petites dimensions pour sélectionner des groupes d'individus homogènes (les classes, les clusters). Ces clusters peuvent être extraits par le processus automatique comme ils peuvent être sélectionnés interactivement par l'utilisateur. Notre approche permet d'évaluer les différents clusters extraits au fur et à mesure, l'utilisateur peut ainsi, être guidé dans le processus d'extraction. Nous proposons également un protocole de test afin de comparer l'approche semi-interactive avec l'approche purement automatique et avec l'approche purement interactive sans aucune aide automatique. Nous verrons que l'approche semi-interactive s'est montrée plus efficace en termes de résultats de clustering.

2 Présentation de l'approche

Nous proposons une approche itérative qui permet d'extraire les classes les unes après les autres. Le processus itératif est répété à la demande de l'utilisateur. A chaque itération, une nouvelle classe est extraite. Il existe plusieurs méthodes pour extraire une classe homogène. Dans cet article, nous utilisons une méthode d'extraction de classes basée sur la détection de limite de classe (Blansché et Boudjeloud-Assala, 2013). Une classe est extraite à partir d'un centre. Nous calculons la distance entre chaque objet et le centre de la classe et cherchons alors la première augmentation abrupte dans ces valeurs qui indiquera la limite de la classe extraite. Nous avons opté pour la méthode de détection de pics présentée dans (Palshikar, 2009), appliquée sur le différentiel des distances. Une fois la limite de classe évaluée, tous les objets qui ont une distance inférieure à cette limite appartiennent à la classe extraite. Concernant l'évaluation des classes, nous proposons d'utiliser deux critères d'évaluation pour évaluer les classes indépendamment les unes des autres, développés par Blasnché et Boudjeloud dans Blansché et Boudjeloud-Assala (2013). Contrairement aux critères d'évaluation en classification non supervisée qui évaluent l'ensemble de données dans son intégralité et ne donnent pas une évaluation des classes indépendamment les unes des autres. Le premier critère est le rapport d'inertie IR (rapport entre l'inertie intra-classe et l'inertie totale des données, normalisé par le nombre d'objets dans la classe et dans l'ensemble de données où C_k représente la k-ième classe extraite :

$$IR\left(C_k\right) = \frac{Card(D) \sum_{o \in C_k} d\left(o, c_k\right)^2}{Card(C_k) \sum_{o \in D} d\left(o, g\right)^2}$$

Le second critère représente le rapport de limite de la classe CLR, représentant le rapport entre la distance du dernier objet de la classe sur la distance du premier objet hors de la classe :

$$CLR\left(C_k\right) = \frac{\max_{o \in C_k} \left(d\left(o, c_k\right)\right)}{\min_{o \notin C_k} \left(d\left(o, c_k\right)\right)}$$

Comme chaque classe est extraite individuellement, nous devons également nous assurer que celles-ci sont différentes les unes des autres. Nous proposons, donc, d'ajouter une péna-

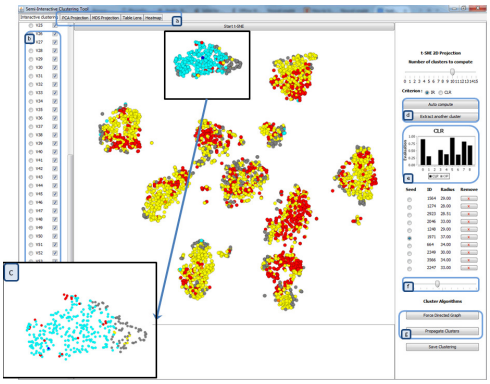

FIG. 1 – *Interface de l'outil semi-interactif.*

lisation des classes selon leur chevauchement avec les classes précédemment découvertes. Le critère de pénalité OP calcule une pénalité selon l'intersection et l'union de la classe extraite avec les classes précédentes ($\lambda \geq 0$ représente le poids de la pénalité selon l'importance que l'on donne aux chevauchements).

$$OP\left(C_k\right) = \lambda \max_{i=1\ldots k_1} \frac{Card(C_k \cap C_i)}{Card(C_k \cup C_i)}$$

Une interface visuelle permettant d'interagir avec le processus d'extraction itératif de classe est également proposée (figure 1). L'ensemble de données est projeté avec t-SNE (van der Maaten et Hinton, 2008). La projection de l'ensemble de données permet à l'utilisateur d'interagir et de sélectionner lui-même des centres de clusters, d'évaluer les clusters sélectionnés, de fixer de manière interactive la limite (rayon) (figure 1-f) et ainsi voir l'effet sur la projection avec les critères d'évaluation décrits précédemment, qui sont mis à jour (figure 1-e) et qui s'affichent au fur et à mesure que les clusters sont construits. Les objets de l'ensemble de données qui n'ont pas été affectés dans les différents clusters par l'approche apparaissent en gris et les objets qui appartiennent à deux classes ou plus apparaissent en rouge (figure 1). Les objets les plus proches des centres, en termes de distances, sont affectés au cluster représenté par son centre et son rayon. La distance est calculée sur l'espace d'origine de l'ensemble de données. L'utilisateur peut ainsi manipuler les données jusqu'à ce qu'il soit satisfait des résultats obtenus et valide ensuite le clustering obtenu en enregistrant les résultats sous formes de classes. Il peut ainsi interagir avec le processus d'extraction et d'exploration ou le laisser travailler indépendamment. Si l'utilisateur n'est pas satisfait il peut sélectionner d'autres centres sur lesquels il peut reproduire le même processus.

3 Scénarios d'usage

Nous allons décrire, dans ce qui suit, les différents scénarios présentés à l'utilisateur. Nous utilisons l'outil développé pour l'exploration et le clustering des données. La variable classe

est supprimée et nous comparons les résultats obtenus avec les vraies classes selon plusieurs critères d'évaluation du clustering (Jaccard, Rand, Rogers-Tanimoto (R-T) et de similarité).

Scénario 1 : Purement interactive Le premier scénario consiste à tester les différentes possibilités d'interaction avec la projection de l'ensemble de données selon les tâches du clustering. Nous voulons obtenir des clusters en sélectionnant les classes de façon purement interactive sans aucune intervention de l'approche automatique(partie droite de l'interface : figure 1-d, e, g).

Scénario 2 : Semi-interactive Le deuxième scénario consiste à tester les différentes possibilités d'interaction visuelle combinées à l'approche automatique. En partant de la même projection des données réalisée avec t-SNE, l'utilisateur peut choisir de lancer l'approche automatique en fixant un nombre a priori de clusters, ou en demandant de les extraire au fur et à mesure par le processus itératif (figure 1-d). En examinant les critères d'évaluation $IR + OP$ et/ou $CLR + OP$ (figure 1-e), l'utilisateur peut également, décider de fixer lui-même les prochains centres et rayons, ou bien, sur chaque centre fixé manuellement, faire varier le rayon (figure 1-f). Les résultats des critères d'évaluation sont mis à jour suite à chaque modification de l'utilisateur.

Scénario 3 : Purement automatique Le dernier scénario consiste à appliquer l'approche purement automatique sur l'ensemble de données, en exécutant le processus itératif d'extraction de classes avec les deux critères d'évaluation et en fixant le nombre de classes au nombre de classes réel.

4 Evaluation utilisateurs

Choix de l'ensemble de données test Nous avons décidé de tester l'approche avec l'ensemble de données OpticalDigits (5 620 objets, 64 attributs et 10 classes réelles) de l'UCI Machine Learning Repository (Blake et Merz (1998)). Il s'agit d'un ensemble de données décrivant la reconnaissance optique de chiffres manuscrits. L'ensemble de données contient une représentation statique de l'image générée à la suite du mouvement de la pointe du stylet. La projection de cet ensemble de données avec l'approche t-SNE est présentée dans la figure 1. Cette projection est obtenue avec les 64 attributs de l'ensemble de données, la distance est également calculée à partir de l'espace d'origine (64 attributs). Notre objectif est d'évaluer les résultats du clustering (automatique, interactif, semi-interactif) avec les classes réelles de l'ensemble de données (les caractères numériques).

Participants Nous avons recruté onze participants âgés de 30 à 40 ans, tous membres de notre laboratoire. Tous les participants sont des chercheurs dans le domaine de l'informatique. Parmi eux deux participants ont des thématiques de recherche dans la visualisation de l'information et de l'interaction en général, trois d'entre eux ont des thématiques de recherche dans la fouille de données et les autres avec des thématiques différentes.

Protocole Nous avons tout d'abord, expliqué l'ensemble de données et l'objectif de notre outil. Nous avons fait un test interactif afin d'expliquer comment l'outil fonctionne et ensuite nous avons expliqué le premier scénario. Cette étape prend cinq minutes, nous avons laissé l'utilisateur faire le premier exercice qui a également duré cinq minutes.

Lorsque l'utilisateur termine son processus de clustering, il enregistre les résultats. Ensuite, nous présentons rapidement le second scénario avec l'approche automatisée, puis on laisse l'utilisateur travailler. Cette étape prend moins de cinq minutes avec l'explication. L'utilisateur devait trouver les clusters les plus significatifs tout en réduisant le chevauchement (points en rouge) et le nombre d'objets non classés (points en gris). Lorsque l'utilisateur termine les deux exercices, nous lui montrons les classes réelles de l'ensemble de données.

5 Résultats

Scénario	Critère	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	Moyenne
Sc 1	Similarité	0.65	0.64	0.54	0.69	0.71	0.59	0.60	0.68	0.66	0.65	0.63	0.64
Sc 2	Similarité	0.66	0.72	0.67	0.64	0.73	0.68	0.71	0.74	0.63	0.69	0.64	**0.68**
Sc 3	Similarité	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63	*0.63*
Sc 1	Rand	0.89	0.86	0.79	0.89	0.90	0.87	0.81	0.90	0.85	0.85	0.85	0.86
Sc 2	Rand	0.86	0.91	0.88	0.85	0.91	0.89	0.85	0.92	0.84	0.89	0.83	**0.88**
Sc 3	Rand	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	*0.79*
Sc 1	Jaccard	0.33	0.30	0.22	0.36	0.39	0.31	0.22	0.38	0.28	0.30	0.28	0.31
Sc 2	Jaccard	0.30	0.38	0.32	0.27	0.42	0.36	0.29	0.43	0.27	0.36	0.25	**0.33**
Sc 3	Jaccard	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	*0.20*
Sc 1	R-T	0.83	0.85	0.82	0.86	0.87	0.84	0.82	0.85	0.84	0.85	0.83	**0.84**
Sc 2	R-T	0.84	0.85	0.84	0.82	0.87	0.85	0.85	0.86	0.84	0.86	0.82	**0.84**
Sc 3	R-T	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	*0.81*
Sc 1	Nbr Cluster	14	11	9	9	**10**	11	**10**	**10**	11	9	15	-
Sc 2	Nbr Cluster	11	15	**10**	13	9	**10**	**10**	12	9	9	12	-
Sc 3	Nbr Cluster	10	10	10	10	10	10	10	10	10	10	10	-

TAB. 1 – *Résultats numériques.*

Après avoir obtenu les deux résultats (scénario 1 et 2, tableau 1) des participants, nous avons réalisé un test de Shapiro-Wilk pour les quatre critères d'évaluation (Jaccard, Rand, Rogers-Tanimoto (R-T) et de similarité). Le tableau 2 résume les résultats obtenus. Les différentes p-value sont supérieures à la valeur de $\alpha = 0.05$, l'hypothèse nulle que les résultats du clustering suivent une loi normale ne peut être rejetée. Nous pouvons donc utiliser le test de Student pour déterminer si les résultats moyens sont, deux à deux, significativement différents. Nous avons comparé : l'approche interactive versus automatique, semi-interactive versus automatique, et enfin, semi-interactive versus interactive. Nous pouvons voir dans le tableau 3 que les approches semi-interactive et interactive sont globalement meilleures que l'approche automatique et que pour le critère de similarité l'approche semi-interactive est meilleure que l'approche interactive.

Approches	Similarité p-value	Rand p-value	Jaccard p-value	R-T p-value
Interactive	0.8134	0.2453	0.5768	0.8752
semi-Interactive	0.5987	0.4556	0.4711	0.7453

TAB. 2 – *Test de Shapiro-Wilk.*

Approches	Similarité p-value	Rand p-value	Jaccard p-value	R-T p-value
Interactive/Automatique	*0.4178*	10^{-5}	10^{-4}	10^{-4}
semi-Interactive/Automatique	0.0011	10^{-6}	10^{-5}	10^{-5}
Semi-Interactive/Interactive	0.0368	*0.2029*	*0.1968*	*0.5804*

TAB. 3 – *Test de Student.*

6 Conclusion

Nous avons présenté dans cet article une approche de clustering visuel semi-interactif avec une évaluation quantitative des résultats utilisateurs. Nous sommes partie de l'hypothèse que la perception visuelle pouvait aider les algorithmes automatiques à obtenir de meilleurs résultats. Pour l'utilisateur, l'approche semi-interactive proposée permet non seulement d'évaluer les clusters sélectionnés visuellement en fonction d'un critère déterminé mais aussi de mesurer comment la sélection d'un nouveau cluster influe sur ceux précédemment extraits. Les résultats statistiques montrent tout d'abord que l'approche semi-interactive obtient de meilleurs résultats en classification non supervisée. Les premiers résultats sont très encourageants et nous poussent à proposer l'évaluation d'une classification faites sur différents sous-espaces (sélection de variables) permettant ainsi d'exploiter la perception visuelle et l'interactivité pour les approches de biclustering et de multi-vues.

Références

Blake, C. et C. Merz (1998). `UCI` repository of machine learning databases. Technical report, University of California, Irvine, Dept. of Information and Computer Sciences. http ://archive.ics.uci.edu/ml/data sets.html. Accedé en janvier 2014.

Blansché, A. et L. Boudjeloud-Assala (2013). Processus itératif d'extraction de classes en non supervisée. In *Extraction et gestion des connaissances (EGC'2013), Actes, 29 janvier - 01 février 2013, Toulouse, France*, pp. 9–14.

Palshikar, G. (2009). Simple algorithms for peak detection in time-series. In *Proceedings of 1st International Conference on Advanced Data Analysis Business Analytics and Intelligence*, pp. 2–13.

van der Maaten, L. et G. Hinton (2008). Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research 9*, 2579–2605.

Summary

This paper proposes a semi-interactive system for visual data exploration using a clustering that combines an automatic approach with an interactive one. The user can manually perform the clustering, he can also choose to let the automated approach find optimal solutions and then interact with the process to improve the clustering results according to his visual perception and domain knowledge. The experiments show that the semi-interactive approach obtains better results than others.