# Proposition d'outil de clustering visuel et interactif

Pierrick Bruneau, Philippe Pinheiro, Bertjan Broeksema, Benoît Otjacques

Centre de Recherche Public Gabriel Lippmann
41, rue du Brill
L-4422 Belvaux
{bruneau | pinheiro | broeksem | otjacque}@lippmann.lu,
http://www.lippmann.lu

**Résumé.** Cet article présente un nouvel outil visuel de clustering interactif. Il utilise une technique de réduction de dimensionnalité pour permettre une représentation 2D des données et des classes associées, initialement établies de manière non-supervisée. L'originalité de l'outil consiste à autoriser des modifications itératives à la fois du clustering et de la projection 2D. Grâce à des contrôles adaptés, l'utilisateur peut ainsi injecter ses préférences, et observer le changement induit en temps réel. La méthode de projection utilisée suit une métaphore physique, qui facilite le suivi des changements par l'utilisateur. Nous montrons un exemple illustrant l'intérêt pratique de l'outil [1].

## 1 Introduction

Dans le contexte d'une fouille exploratoire, le recours à des techniques de réduction de dimensionnalité permet classiquement de contourner la difficulté de représenter des résultats de clustering réalisés sur des données à haute dimensionnalité (HD). Les étiquettes de clusters, associées par exemple à des couleurs catégorielles, peuvent alors être appliquées aux points d'un nuage 2D ou 3D.

Les techniques de réduction de dimensionnalité sont susceptibles d'introduire des artéfacts de déchirement et de recollement (Aupetit, 2007). Les algorithmes de clustering ne sont pas sujets à ces artéfacts, mais peuvent mener à des résultats sous-optimaux, ou avoir été mal paramétrés. L'objet de cet article est de proposer un outil interactif de fouille visuelle combinant le meilleur de ces deux approches. Il utilise une projection 2D obtenue par t-SNE, une technique de réduction de dimensionnalité non-supervisée (van der Maaten et Hinton, 2008). Nous ne proposons pas un algorithme de clustering *per se*, mais plutôt une manière itérative d'améliorer conjointement un clustering initial calculé de manière non-supervisée dans un espace HD, et une représentation 2D associée.

Une présentation générale de notre outil est proposée en section 2. Les clusters sont amendés grâce à des techniques de diffusion d'étiquettes, présentées en section 3. Réciproquement, l'adaptation de la projection 2D aux clusters est évoquée dans la section 4. Les exemples donnés en section 5 et tout au long de cet article utilisent le jeu de données COIL-20 (Nene et al.,

1. Cet article est un résumé de *Cluster Sculptor, an interactive visual clustering system, Neurocomputing 150-B 627-644, 2015*, des mêmes auteurs.

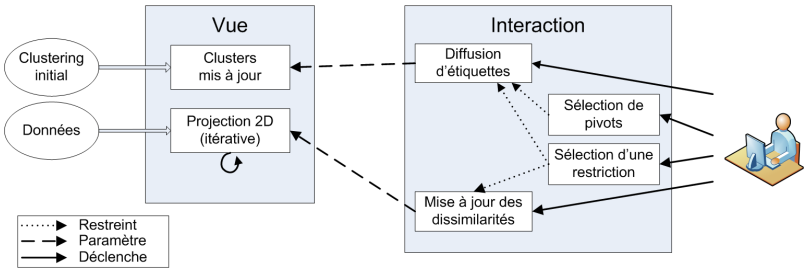


FIG. 1 – *Diagramme résumant la logique de l'outil.*

1996). Il contient 1440 images, réparties en 10 classes. Les images sont décrites par les intensités de leurs 1024 pixels sur une échelle de gris.

## 2 Présentation de l'outil

La logique de l'outil est résumée dans la figure 1. Il est paramétré par un jeu de données HD (i.e., $> 3$), et par un clustering non-supervisé réalisé dans l'espace HD.

Une projection 2D initiale est calculée de manière non-supervisée par l'algorithme itératif t-SNE. Elle est matérialisée par un nuage de points, dont les colorations sont associées aux étiquettes de clusters via un ensemble de couleurs catégorielles. À partir d'une projection et d'un clustering donnés, l'utilisateur peut déclencher les actions suivantes :

— *Sélection d'une restriction* : optionnellement, l'utilisateur peut limiter le rayon de son action à un ensemble de clusters sélectionnés directement en cliquant la légende (voir Figure 2e).

— *Sélection de pivots* : en cliquant sur un des points dans le nuage (voir Figure 2d), l'utilisateur définit un pivot pour la diffusion d'une nouvelle étiquette de cluster. Un inspecteur interactif est à sa disposition pour associer une sémantique à chaque pivot (voir Figure 2c).

— *Diffusion d'étiquettes* : les étiquettes des pivots sélectionnés sont diffusées en utilisant la proximité 2D entre éléments.

— *Modification des dissimilarités* : la répartition en clusters peut être utilisée pour influencer les dissimilarités sous-jacentes à l'algorithme t-SNE. L'utilisateur peut paramétrer le niveau de cet impact.

Les actions de l'utilisateur modifient la visualisation et les clusters, dont le nouvel état peut servir d'entrée à une nouvelle itération du diagramme en figure 1. Des itérations sont effectuées jusqu'à ce que l'utilisateur soit satisfait du résultat.

## 3 Diffusion d'étiquettes

Supposons tout d'abord que l'utilisateur souhaite amender les clusters en utilisant la répartition 2D des éléments. Pour faciliter le processus de modification, et s'éviter des sélections

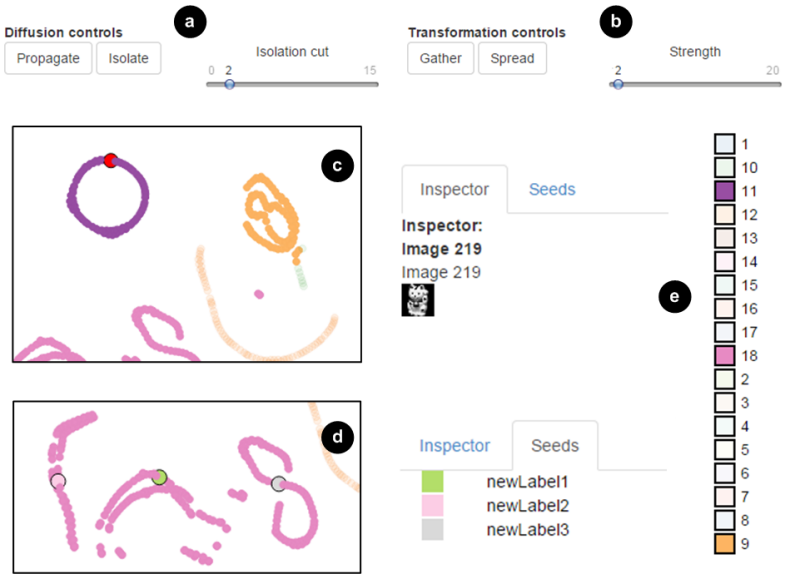


FIG. 2 – a) *Contrôles de diffusion d'étiquettes.* b) *Contrôles de modification des dissimilarités.* c) *Le survol des éléments déclenche un inspecteur interactif. Les éléments hors de la restriction en cours sont affichés en transparence.* d) *Les pivots sélectionnés sont listés dans un panneau éditable.* e) *Le contrôle de la restriction se fait par une légende cliquable.*

fastidieuses, nous proposons une diffusion semi-automatique basée sur la sélection de pivots par l'utilisateur. La diffusion peut être calculée selon deux algorithmes issus de la littérature :

— *Propagation probabiliste :* les étiquettes peuvent métaphoriquement sauter de manière probabiliste depuis les pivots associés, puis d'élément en élément. Ce processus converge, et le résultat peut être obtenu sous une forme analytique impliquant de simples produits de matrices (Zhu et Ghahramani, 2002).
— *Coupes de l'arbre de couverture minimal :* l'arbre de couverture minimal d'un graphe peut être calculé grâce à l'algorithme de Kruskal (Kruskal, 1956). Des coupes dans cet arbre isolent des composantes connexes du graphe.

Ces opérations sont réalisées relativement à la distribution visuelle des éléments ; les distances 2D entre éléments dans la projection sont donc utilisées dans les algorithmes.

L'utilisateur commence par sélectionner un ou plusieurs pivots dans la restriction en cours (qui peut englober tous les éléments si aucun cluster n'a été sélectionné dans la légende). Ces derniers peuvent être vus comme des prototypes de clusters, existants et à redécouper, ou à créer. Selon ses préférences, l'utilisateur peut alors les propager exhaustivement, ou paramétrer interactivement la coupe de l'arbre de couverture minimal pour isoler des composantes connexes. Dans ce dernier cas, il peut aussi choisir de regrouper les composantes sans pivot dans un cluster résiduel, ou laisser leur étiquette telle qu'avant l'interaction.

# 4 Modification des dissimilarités

Plutôt que d'utiliser la distribution 2D des éléments pour modifier les clusters comme dans la section 3, l'utilisateur peut utiliser la répartition en clusters pour modifier la projection 2D, de manière par exemple à renforcer la séparation des clusters.

Considérons le graphe complet entre les éléments dans la restriction en cours, pondéré par les dissimilarités dans l'espace HD. Nous voulons utiliser l'information portée par les clusters pour amender les dissimilarités HD entre éléments. Pour préserver la structure interne des clusters, nous proposons de restreindre la modification au sous-graphe multipartite induit par les clusters. La fonction cumulative normalisée de la distribution Beta est alors appliquée aux poids d'arêtes associés :

$$P^{[a,b]}_{\text{beta}(\alpha,\beta)}(x) = (b-a)P^{[0,1]}_{\text{beta}(\alpha,\beta)}\left(\frac{x-a}{b-a}\right) + a \tag{1}$$

Les bornes $a$ et $b$ permettent d'adapter la transformation aux valeurs de dissimilarité, e.g., à la valeur maximale observée dans la restriction, ou à la cohésion interne des clusters. Les changements trop disruptifs sont ainsi évités. L'utilisateur peut alors paramétrer interactivement un rapprochement (respectivement un éloignement) des clusters en augmentant le paramètre $\alpha$ (respectivement $\beta$).

L'algorithme t-SNE se base sur des dissimilarités HD entre les éléments pour estimer leurs positions dans le nuage de points. Classiquement, ces dissimilarités sont initialisées par une distance Euclidienne dans l'espace HD. Dans l'outil, les dissimilarités peuvent être modifiées dynamiquement.

L'algorithme t-SNE peut être interprété comme une variante d'algorithme force et ressort. La métaphore physique suivie par cette classe d'algorithmes convertit des changements discontinus des forces en présence en mouvements continus. Ainsi, la discontinuité obtenue à l'application de l'équation (1) est convertie en mouvements continus, facilitant leur suivi par un utilisateur. Après une modification de dissimilarités, ce dernier peut alors suivre le changement progressif induit par son action.

Les dissimilarités HD demeurent latentes à la projection 2D, et ne peuvent pas être observées directement dans la visualisation. Pour pallier cette limitation, nous avons incorporé l'outil ProxiViz (Heulot et al., 2012), qui permet de mapper interactivement les dissimilarités sur le diagramme de Voronoï du nuage de points (voir Figure 3). L'utilisateur dispose ainsi d'une information plus complète avant de procéder à ses modifications. Ceci peut par exemple permettre de prendre en compte d'éventuels artéfacts de projection (Aupetit, 2007).

# 5 Exemples

Au cours de ses manipulations avec l'outil, l'utilisateur est confronté à la situation de la figure 4a. Le cluster vert, identifié par l'algorithme de clustering dans l'espace HD, est éclaté en 3 composantes dans la visualisation. L'utilisateur souhaite les regrouper, en respectant le voisinage des composantes dans la projection.

Il commence par vérifier la pertinence d'un tel regroupement en utilisant l'outil ProxiViz. Les composantes du cluster vert sont bruitées par les clusters violet et orange. Il commence

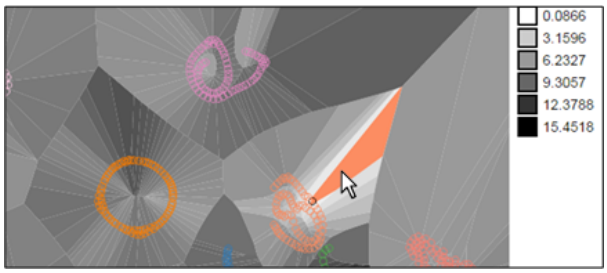


FIG. 3 – *Le survol du nuage de points déclenche l'outil ProxiViz. Les dissimilarités HD par rapport au point de la cellule survolée sont mappées sur une échelle de gris, et colorent les cellules de Voronoï des éléments respectifs. Les étiquettes de clusters sont rappelées en colorant le contour des points, ainsi que la cellule du point survolé.*

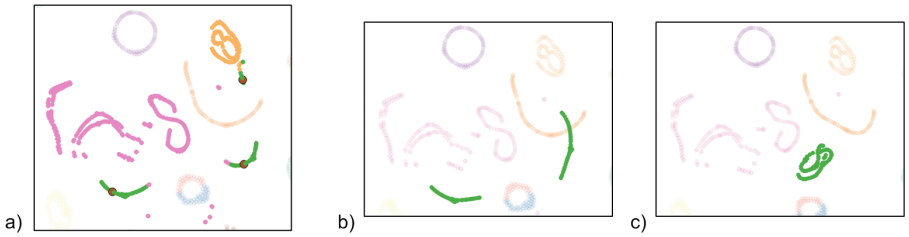


FIG. 4 – a) *Le cluster vert est réparti en 3 composantes, bruité par les clusters violet et orange. Un pivot est sélectionné pour chaque composante.* b) *L'exécution d'itérations de t-SNE sur la restriction en cours ne réunit que partiellement le cluster.* c) *Après mise à jour des dissimilarités, le cluster est effectivement regroupé.*

donc par définir une restriction à ces 3 clusters, et sélectionne des pivots pour isoler les composantes (voir Section 3). Les composantes sont ensuite réunies simplement en éditant la légende interactive.

En déclenchant t-SNE sur la restriction en cours, le cluster est partiellement réuni (voir Figure 4b). L'utilisateur influence leur rapprochement en mettant à jour les dissimilarités (voir Section 4 et Figure 4c).

# 6 Conclusion

Dans cet article, nous avons exposé notre outil de clustering visuel et interactif. Les techniques de diffusion d'étiquettes et de modification des dissimilarités permettent d'enrichir mutuellement une projection 2D et un clustering itérativement mis à jour. La métaphore physique suivie par le nuage de points et les moyens de contrôle offerts par l'interface permettent à l'utilisateur de suivre le changement progressif induit par ses actions. Nous avons illustré l'intérêt de l'approche au travers d'exemples.

De nombreuses perspectives s'ouvrent pour enrichir et améliorer cet outil. Outre des fonctionnalités génériques comme la tenue d'un historique des interactions, les techniques proposées (e.g., sélection de pivots, propagation probabiliste, isolement de composantes) mériteraient d'être testées indépendamment. Les différences et complémentarités éventuelles avec des travaux existants (Lee et al., 2012), parfois affiliés au domaine de l'apprentissage de métrique (Brown et al., 2012; Martin et al., 2012), doivent également être étudiées.

# Références

Aupetit, M. (2007). Visualizing distortions and recovering topology in continuous projection techniques. *Neurocomputing 70*(7-9), 1304–1330.

Brown, E. T., J. Liu, C. E. Brodley, et R. Chang (2012). Dis-function : Learning distance functions interactively. In *IEEE Conference on Visual Analytics Science and Technology*, pp. 83–92.

Heulot, N., M. Aupetit, et J.-D. Fekete (2012). Proxiviz : an interactive visualization technique to overcome multidimensional scaling artifacts. In *Proceedings of IEEE InfoVis, poster*.

Kruskal, J. B. (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. In *Proceedings of the American Mathematical Society*, Volume 7, pp. 48–50.

Lee, H., J. Kihm, J. Choo, J. Stasko, et H. Park (2012). iVisClustering : An interactive visual document clustering via topic modeling. In *Computer Graphics Forum*, Volume 31, pp. 1155–1164. Wiley Online Library.

Martin, L., M. Exbrayat, G. Cleuziou, et F. Moal (2012). Interactive and progressive constraint definition for dimensionality reduction and visualization. *Advances in Knowledge Discovery and Management, Studies in Computational Intelligence 398*, 121–136.

Nene, S. A., S. K. Nayar, et H. Murase (1996). Columbia object image library (coil-20). Technical report, Technical Report CUCS-005-96.

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

Zhu, X. et Z. Ghahramani (2002). Learning from labeled and unlabeled data with label propagation. Technical report, Technical Report CMU-CALD-02-107, Carnegie Mellon University.

# Summary

This paper introduces a novel visual and interactive clustering tool. It relies on a dimensionality reduction technique to allow for a 2D representation of the data and associated clustering, set initially in an unsupervised fashion. Its main contribution is to enable iterative updates of both the 2D projection and clustering. Using appropriate controls, the user may thus inject his or her preferences, and visualize the induced change in real time. The involved dimensionality reduction technique follows a physical metaphor, that facilitates the tracking of changes by the user. The practical interest of the tool is illustrated by an example.