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.