Clustrophile 2: Guided Visual Clustering Analysis

Marco Cavallo
IBM Research
mcavall@us.ibm.com

Çağatay Demiralp
IBM Research
cagatay@cs.stanford.edu

ABSTRACT

Data clustering is a common unsupervised learning method frequently used in exploratory data analysis. However, identifying relevant structures in unlabeled, high-dimensional data is nontrivial, requiring iterative experimentation with clustering parameters as well as data features and instances. The space of possible clusterings for a typical dataset is vast, and navigating in this vast space is also challenging. The absence of ground-truth labels makes it impossible to define an optimal solution, thus requiring user judgment to establish what can be considered a satisfiable clustering result. Data scientists need adequate interactive tools to effectively explore and navigate the large space of clusterings so as to improve the effectiveness of exploratory clustering analysis.

We introduce Clustrophile 2, a new interactive tool for guided clustering analysis. Clustrophile 2 guides users in clustering-based exploratory analysis, adapts user feedback to improve user guidance, facilitates the interpretation of clusters, and helps quickly reason about differences between clusterings. To this end, Clustrophile 2 contributes a novel feature, the clustering tour, to help users choose clustering parameters and assess the quality of different clustering results in relation to current analysis goals and user expectations.

We evaluate Clustrophile 2 through a user study with 12 data scientists, who used our tool to explore and interpret sub-cohorts in a dataset of Parkinson's disease patients. Results suggest that Clustrophile 2 improves the speed and effectiveness of exploratory clustering analysis for both experts and non-experts.

KEYWORDS

Exploratory data analysis, visual analytics, clustering, unsupervised learning, dimensionality reduction, Clustering Tour, guidance, interpretability

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

IDEA @ KDD'18, August 20th, 2018, London, United Kingdom

© 2018 Copyright held by the owner/author(s).

ACM Reference Format:
Marco Cavallo and Çağatay Demiralp. 2018. Clustrophile 2: Guided Visual Clustering Analysis. In *Proceedings of KDD 2018 Workshop on Interactive Data Exploration and Analytics (IDEA'18) (IDEA @ KDD'18).* ACM, New York, NY, USA, 1 page.

This paper has been accepted for publication at IEEE Vis VAST 2018. The original draft can be accessed at: https://arxiv.org/abs/1804.03048