Interactive Unsupervised Clustering with Clustervision

Work-in-Progress

Bum Chul Kwon
IBM T.J. Watson Research Center
bumchul.kwon@us.ibm.com

Ben Eysenbach
Massachusetts Institute of Technology
bce@mit.edu

Janu Verma
IBM T.J. Watson Research Center
jverma@us.ibm.com

Kenney Ng
IBM T.J. Watson Research Center
kenney.ng@us.ibm.com

Adam Perer
IBM T.J. Watson Research Center
adam.perer@us.ibm.com

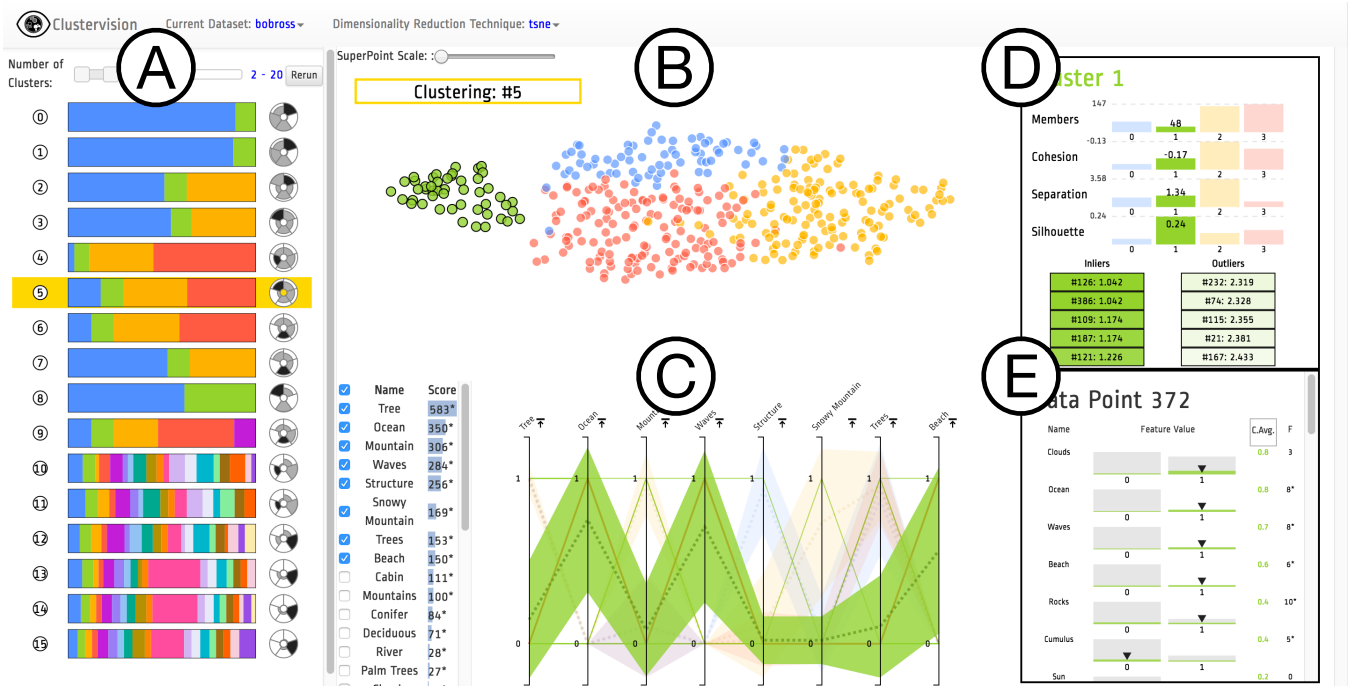

Figure 1: An overview of *Clustervision* on a dataset describing 403 paintings by the "Joy of Painting" artist Bob Ross. (A) *Ranked List of Clustering Results* shows 16 different clustering results that are sorted by the aggregated quality measures; (B) *Projection* shows a selected clustering result (highlighted in yellow in (A)) on a projection of data points colored corresponding to corresponding clusters; (C) *Parallel Trends* show the trends of feature values of data points within corresponding clusters in areas across parallel coordinates. Cluster 1 (Green Color) is highlighted; (D) *Cluster Detail* shows quality measures of a selected individual cluster (Cluster 1); (E) *Data Point* shows the feature value distribution of the selected cluster as well as the selected data point (Data Point 372 within Cluster 1).

KEYWORDS

Unsupervised Clustering, Visual Analytics, Quality Metrics, Interactive Visual Clustering

ACM Reference format:
Bum Chul Kwon, Ben Eysenbach, Janu Verma, Kenney Ng, and Adam Perer. 2017. Interactive Unsupervised Clustering with Clustervision. In *Proceedings of IDEA Workshop Submission, Halifax, NS, Canada, 2017 (IDEA'17),* 4 pages.
https://doi.org/10.475/123_4

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'17, 2017, Halifax, NS, Canada

© 2017 Copyright held by the owner/author(s).

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

1 INTRODUCTION

Clustering, the process of grouping together similar items into distinct partitions, is a common type of unsupervised machine learning that can be useful for summarizing and aggregating complex multi-dimensional data. However, data can be clustered in many ways, and

there exist a large number of algorithms designed to reveal different patterns. While having access to a wide variety of algorithms is helpful, in practice, it is quite difficult for data scientists to choose and parameterize algorithms to get the clustering results relevant for their dataset and analytical tasks. To alleviate this problem, we built a clustering analysis system, *Clustervision*, that helps ensure data scientists find the right clustering among the large amount of techniques and parameters available. Our system clusters data using a variety of clustering techniques and parameters and then recommends good clustering results utilizing a variety of quality scoring metrics. In addition, users can guide the system to produce more relevant results by providing task-relevant constraints on the data. Our visualization interface allows users to find high quality clustering results, explore the clusters using a variety of coordinated visualization techniques, and select the cluster result that best suits their task.

2 CLUSTERVISION

In order to support interactive exploration of clustering results, we propose *Clustervision*. In this paper, we demonstrate the system using a small but illustrative dataset of 403 paintings produced on the PBS show "The Joy of Painting". Over the course of the 403 episodes, a variety of diverse landscapes were painted, which were manually coded by FiveThirtyEight[1] using 67 features (e.g. trees, water, mountains, and weather elements).

After a dataset is loaded into the tool, *Clustervision* computes and evaluates all possible combinations of clustering techniques and parameters. In this configuration, *Clustervision* will use three clustering techniques (k-means, Spectral Clustering, and Agglomerative Clustering) and 19 parameter configurations (k=2-20), resulting in 58 clustering results. The system can also optionally include more clustering techniques and parameters. Each of the clustering results are then analyzed using, by default, 5 quality metrics (Calinski-Harabaz[8], Silhouette[11], Davies-Bouldin[5], S_{Dbw}[7], and Gap Statistic[13]). As each of these quality metrics aim to compute quality using different properties of the clusters (e.g. variance, within-cluster distance, between-cluster distance, density), we chose not to rely on a single metric but instead a variety of diverse metrics. By default, the top 3 highest ranking results from each metric are presented to the user, resulting in 15 results top results for the user to consider. In order to ensure the results aren't too similar, an item will only be considered as a top result if its at least 5% different from another top result.

Figure 1(a) shows an example of the top 15 clustering results. Each row features a clustering summary glyph, where each colored stripe represents a color whose width is proportional to the number of data points in that cluster. Each cluster has a unique color that is consistently used across all views in the UI. On the right is a quality summary glyph that shows values of each of the five quality metrics. Similar to a pie chart, the glyph is a circular region divided into five equal slices for each of the metrics.

In order to understand if a particular clustering result is relevant to the analytical task, users often need to see their data points in context of the cluster groupings. The *Projection* view encodes data points as circular elements in a two dimensional space, resembling a scatterplot, as shown in Figure 1(b). However, instead of plotting

[1] https://fivethirtyeight.com/features/a-statistical-analysis-of-the-work-of-bob-ross/

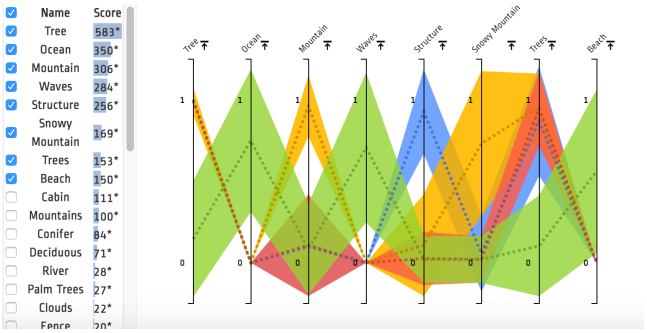

Figure 2: The *Parallel Trends* view is similar to parallel coordinates, but in order to simplify the complexity of many lines, the view focuses on showing the trends of each cluster. *Parallel Trends* has vertical axes that represents each feature of the data points. However, instead of drawing a line crossing the axes for each data point as in parallel coordinates, *Parallel Trends* draws an area path per cluster. The intervals cross each axes, where the vertical ends represent standard deviation or 95% confidence intervals for the corresponding features.

the data on only two dimensions of the data, *Clustervision* uses dimensionality reduction techniques (e.g. t-SNE [10]) to synthesize all of the dimensions. The main use of the *Projection* view is to have a consistent and stable representation, as the positions of the data points remain stable across all clustering results. Although the position of the data points gives clues to the distance and separation between clusters, users get more evidence about the underlying properties of the clusters from the other views. The *Projection* view serves as one way to explore both individual data points and clusters. Most importantly, it allows users to use other views to get more details about the selected data points and clusters.

In order to help summarize the clustering results, the *Ranked Features* and *Parallel Trends* views are coordinated with the projection view and shows information about the features of the selected clustering result. One of the challenges associated with unsupervised clustering is that even after clusters are defined by a technique, it is difficult to summarize why the cluster groupings were made. In an attempt to retrieve the features responsible for the separation, we utilize univariate statistics to compute whether there is a statistically significant relationship between each feature and each cluster. We consider this a classification task, where each cluster is a class, and compute the analysis of variance (ANOVA) for the entire dataset. The result scores based on the ANOVA F-Value allow us to rank each feature in order of importance. These important features are displayed as a ranked list in the *Ranked Features* view, where each feature name is augmented with a numeric importance score and a corresponding bar chart, as shown in Figure 1(c).

The *Parallel Trends* is similar to parallel coordinates, but in order to simplify the complexity of many lines, initially the view only shows the trends of each cluster. As in parallel coordinates, *Parallel Trends* has vertical axes that represents each feature of the data points. However, instead of drawing a line crossing the axes for each data point as in parallel coordinates, *Parallel Trends* draws an area path per cluster. The intervals cross each axes, where the vertical

ends represent standard deviation or 95% confidence intervals for the corresponding features. Then, a dotted line is drawn on top of the area path per cluster to show the mean values for each cluster for the corresponding data feature. To see details of a cluster, users can click on an area path to show individual lines that represent corresponding data points within the cluster as shown in Figure 1(c). This implementation also allows users to sort axes, switch axes, and filter on specific feature values on each axis, which are interaction techniques common to parallel coordinates.

For example, in the selected Bob Ross clustering shown in Figure 2, the top features most responsible for the cluster grouping are the presence of trees, mountains, and oceans in paintings. This ranked list in conjunction with the *Parallel Trends* views help show how these features correlate with the clusters. The Green cluster has uniquely high values in Ocean, Waves, and Beach, giving a clear indication that this cluster represents the ocean-oriented paintings of Ross. This cluster is demonstrably different from the Yellow cluster (which has high values of tree, mountain, snowy mountains, and trees), the Blue cluster (with Structures), and the Red cluster (with tree and trees). While only the top 8 features are shown, other features can be added by selecting them.

The *Cluster Detail* view appears when users select a particular cluster from the *Projection* or *Parallel Trends* views. This view is designed to present a summary of the clusters using statistics and prototypes. For the selected cluster, the number of data points that are members of the cluster is shown as a labeled bar that is the same color of the cluster. This number is put in context with all of the other cluster sizes by showing translucent bars representing each cluster to form a bar chart. Similar bar charts are shown for statistics summarizing the cluster, such as cohesion, separation, and silhouette scores, as shown at the top of Figure 1(d). In addition to these statistical summaries, the *Cluster Detail* view also shows members of the cluster that typical or atypical for the cluster based on the distance metric, as inliers and outliers.

The *Data Point* view appears when users select or mouseover a data point in the *Projection* or *Parallel Trends* views. The *Data Point* provides details about the actual values of a data points features. However, this view also puts them in the context of other other data points by presenting the distribution of values alongside each value. For binary variables and categorical feature values with less than five levels, we show histogram rather than density plot and provide triangle marks to show the selected data point as seen in Figure 1(e).

Users can sort features by their name, value, cluster average value, and importance. The importance calculation is similar to the technique used in the *Ranked Features* view. However, here the technique considers assigns the selected cluster as a first class, and all other clusters as a second class. By computing an ANOVA using these cluster-centric classes, it is possible to determine which features are responsible for why the selected cluster is different from all other clusters. This option presents the most important features at the top of this view, making it easy to compare between data points and clusters by mouse-overing regions of the interest in the *Projection* view.

Users can also interactively request new results by setting up constraints with respect to specific data points. Users can select multiple data points and tell the system that they need to be either in the same cluster or in separate clusters. Then, the system filters clustering results based on the requirements set.

3 RELATED WORK

There have been many previous visualization systems that attempt to employ clustering to support high dimensional data analysis. Hierarchical Clustering Explorer [12] allows users to investigate an overview of a clustering result and to compare details of clusters by using coordinated displays. VISTA [4] enables users to visually view clusters of a clustering result on 2D projection and apply internal quality metric scores. Dicon [3] visualizes multidimensional clusters' quality as well as attribute-based information through icon-based visualization. Unlike *Clustervision*, these systems do not support comparison between multiple clustering results.

Some systems allow users to provide feedback on clustering results so that the next run applies the inputs. desJardins et al. [6] proposed a technique to iteratively run and visualize constrained clustering with constraints made by users. iVisClustering allows users to adjust cluster hierarchy and to re-label individual data items (i.e., documents) into another cluster [9]. Cluster Sculptor also allows users to update cluster labels on a 2D projection [2]. Boudjeloud-Assala *et al.* proposes an interactive visual clustering system that allows users to define seeds and limits of clusters for steering the clustering process [1]. While these systems help steer the user toward better clustering results, the user must define how to make the clustering better rather than receiving recommendations from the system, unlike *Clustervision*.

4 CONCLUSION

In this paper, we described the features of *Clustervision*, a work-in-progress which we believe is a promising interface to help data scientists find meaningful clusterings of their data. By integrating clustering techniques and quality metrics with coordinated visualizations, the system allows users to interactively explore and analyze clustering results at various levels. Although we demonstrated this system on a small dataset in this paper, we are currently deploying this system with a team of data scientists to find meaningful clusters of patients that share complex diseases, which they plan to publish in an upcoming medical journal.

REFERENCES

[1] Lydia Boudjeloud-Assala, Philippe Pinheiro, Alexandre BlanschÃl', Thomas Tamisier, and BenoÃőt Otjacques. 2016. Interactive and iterative visual clustering. *Information Visualization* 15, 3 (July 2016), 181–197.

[2] P. Bruneau, P. Pinheiro, B. Broeksema, and B. Otjacques. 2015. Cluster Sculptor, an interactive visual clustering system. *Neurocomputing* 150, Part B (Feb. 2015), 627–644.

[3] N. Cao, D. Gotz, J. Sun, and H. Qu. 2011. DICON: Interactive Visual Analysis of Multidimensional Clusters. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (Dec. 2011), 2581–2590.

[4] Keke Chen and Ling Liu. 2004. VISTA: Validating and Refining Clusters Via Visualization. *Information Visualization* 3, 4 (Dec. 2004), 257–270.

[5] David L. Davies and Donald W. Bouldin. 1979. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.* 1, 2 (Feb. 1979), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

[6] Marie desJardins, James MacGlashan, and Julia Ferraioli. 2007. Interactive Visual Clustering. In *Proceedings of the 12th International Conference on Intelligent User Interfaces (IUI '07)*. ACM, New York, NY, USA, 361–364.

[7] Maria Halkidi and Michalis Vazirgiannis. 2001. Clustering Validity Assessment: Finding the Optimal Partitioning of a Data Set. In *Proceedings of the 2001 IEEE*

International Conference on Data Mining (ICDM '01). IEEE Computer Society, Washington, DC, USA, 187–194. http://dl.acm.org/citation.cfm?id=645496.657864

[8] Marcin Kozak. 2012. âĂIJA Dendrite Method for Cluster AnalysisâĂİ by CaliÂĎski and Harabasz: A Classical Work that is Far Too Often Incorrectly Cited. *Communications in Statistics - Theory and Methods* 41, 12 (2012), 2279–2280. https://doi.org/10.1080/03610926.2011.560741 arXiv:http://dx.doi.org/10.1080/03610926.2011.560741

[9] Hanseung Lee, Jaeyeon Kihm, Jaegul Choo, John Stasko, and Haesun Park. 2012. iVisClustering: An Interactive Visual Document Clustering via Topic Modeling. *Computer Graphics Forum* 31, 3pt3 (June 2012), 1155–1164.

[10] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, Nov (2008), 2579–2605.

[11] Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1987), 53 – 65. https://doi.org/10.1016/0377-0427(87)90125-7

[12] Jinwook Seo and B. Shneiderman. 2002. *Interactively exploring hierarchical clustering results*.

[13] Robert Tibshirani, Guenther Walther, and Trevor Hastie. 2000. Estimating the number of clusters in a dataset via the Gap statistic. 63 (2000), 411–423.