Eurographics Conference on Visualization (EuroVis), Posters Track (2017)
A. Puig Puig and T. Isenberg (Editors)

DaaG: Visual Analytics Clustering using network representation

Daniel Alcaide[†1,2] and Jan Aerts[1,2]

[1]VDA-lab, ESAT/STADIUS, KU Leuven, Kasteelpark Arenberg 10, 3001 Leuven
[2]IMEC Smart Applications and Innovation Services, Kasteelpark Arenberg 10, 3001 Leuven

Abstract

Finding useful patterns in datasets has attracted considerable interest in the field of visual analytics. One of the most common solutions is the identification and representation of clusters. In this work, we propose a visual analytics clustering methodology for guiding the user in the exploration and detection of clusters in a dataset. We thereby combine the homological algebra with a graphical representation of the clustered dataset as a network into one coherent framework. Our approach entails displaying the results of the heuristics to users, providing a setting from which to start the exploration and data analysis.

Categories and Subject Descriptors (according to ACM CCS): H.3.3 [Information Search and Retrieval]: Clustering—

1. Introduction

Visualizing clustering results can help to quickly assimilate the information and provide insights that support and complement textual outputs or statistical summaries. For example, we quickly wish to know how well defined the clusters are, how different they are from each other, what their size is, and if the observations belong strongly to the cluster or only marginally. Visualizing a clustering solution has many potential uses. The analyst during the iterative threshold selection process can quickly obtain insights from the visualization that suggest the adequacy of the solution and what further experiments to conduct. Moreover, the user can examine and query the final clustering solution using the visualization allowing a better understanding of the original dataset. Alternatively, determining the number of clusters in a dataset is a frequent problem in data clustering, and is a distinct issue from the process of actually solving the clustering problem [Dav02]. The correct choice of the number of groups is often ambiguous, with interpretations depending on the shape and scale of the distribution of points in a dataset and the desired clustering resolution by the user. The optimal choice of clusters depends on the intended use, but in general, it strikes a balance between the maximum compression using a single cluster and the highest resolution of the data by assigning each data point to its own cluster.

Several clustering algorithms have been proposed for partitioning datasets. Most of them need setting parameters for these algorithms, such as the number of cluster in k-means [Mac67], the reference value (epsilon) in DBSCAN [EKSX96] or the cutoff distance in the hierarchical clustering [Cha05]. These parameters vary from one algorithm to another, but most clustering algorithms require a parameter that either directly or indirectly specifies the number of clusters. Setting these parameters demand either detailed pre-existing knowledge of the data or time-consuming trial and error. This latter case needs repeated runs with different parameters and still need that the user has sufficient domain knowledge to know what a good clustering "looks" like. Our approach tries to enhance the interaction between the algorithm and the human throughout visual analytics [SC04].

We suggest a novel and generic approach called DaaG (Distance matrix as a Graph) for grouping and visualizing multiple granularities of the data that enables: (1) exploration of the dataset using a focus-plus-context representation, (2) simplification of the dataset using aggregation based on the similarity of data elements, (3) acceleration of the threshold selection through of homological algebra, and (4) inclusion of the human in the process of selection the number of clusters.

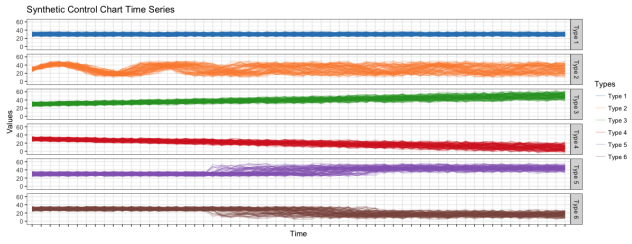

Figure 1: *Representation of the dataset that contains 600 examples of control charts synthetically generated by the process in Alcock and Manolopoulos (1999).*

† Corresponding author: daniel.alcaide@kuleuven.be

DOI: 10.2312/eurp.20171169

2. Method

The developed method consists of two parts; the first is the transformation of the distance matrix into a simplified network and the second is the exploration of the resulting networks for different threshold values. The methodology is illustrated using a dataset taken from UCI repository website [BL13] (Fig. 1).

2.1. Transformation of the distance matrix into a network

The resulting graphs represent the abstraction of the clustering process (Fig. 2 and Fig. 3). All possible subgraphs in which all the vertices are directly or indirectly connected, i.e., all nodes in this subgraph can be reached from any node, represent a cluster. The process to build the network is summarized in the following steps:

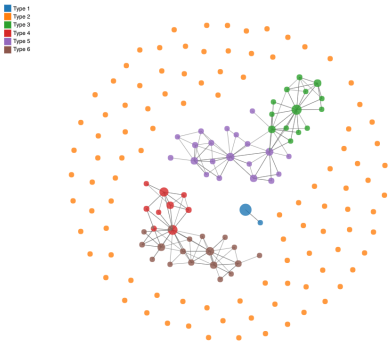

Figure 2: *Visualization of the clustered dataset using a distance threshold of 185. Using this threshold, patterns 3 and 5 appear linked, and patterns 4 and 6 are linked. However, pattern 2 is disconnected due to the noisiness.*

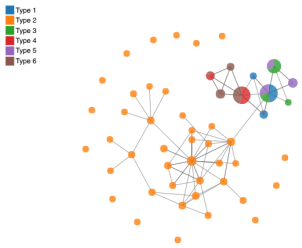

Figure 3: *Visualization of the clustered dataset using a distance threshold of 252.*

1. Once a distance threshold between points is defined, every data element from the dataset is represented as a node in the network.
2. Every pair nodes (i,j) are connected if the distance between nodes i and j is below the threshold established.
3. The node density is defined as the number of connections that the node has in a network. This value is computed for all nodes, and the densest one is the first candidate to be the center of an aggregated node. All nodes connected directly with the candidate are converted into aggregated node (that can include multiple data elements). The rest of the data elements still connected indirectly with the candidates remain connected with the aggregated node.
4. Step number 3 is repeated iteratively with the subsequent candidates until obtaining an aggregated network.

2.2. Stability of clustering

For the proposed methodology, we provide feedback on output stability given a parameter choice. This method involves graphing the number of clusters on the x-axis and the cutoff selected on the y-axis (Fig. 4). A marked slope of the graph suggests that the clusters being combined are very dissimilar. Thus, the appropriate number of clusters is found at the “elbow” of the graph. Interpreting a graph, however, may be difficult; for example, the elbow may not be pronounced, indicating that there may not be any natural groups in the data. Alternatively, the graph may have more than one elbow, indicating that more than one natural set of clusters fit the data [KJS96]. The mechanism to find the optimal threshold is similar to other clustering algorithms like k-means but scales better with respect to the time that it takes to complete each iteration due to the use of algebraic topology to detect and assign the data element into the clusters.

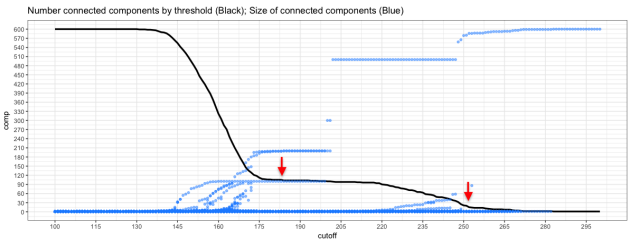

Figure 4: *Number of clusters vs. threshold distance. The "elbows" are indicated by the red arrow. The first one is set at threshold 185 (Fig. 2) and the second one at 252 (Fig. 3). Blue points represent the number of data elements by connected component. For example, when the threshold is 295 there is only one connected component (one single cluster) that includes all the data elements of the dataset.*

3. Conclusion

The clustering problem is challenging because there is a lack of guidance in forming the clusters. One of the main elements that might be of use defining the clusters is the inclusion of knowledge provided by a domain expert. With DaaG, we facilitate the incorporation of the user knowledge by improving the representation of clustered dataset especially in the presence of noise in the data or outliers. DaaG is a work in progress, and future work will include refinement of the detection of the optimal threshold for the current algorithm, the inclusion of additional topological measures and the improvement of the definition of the simplified nodes by establishing a weighted metric. Overall, we believe that DaaG methodology can potentially be an alternative clustering algorithm supporting analysts to deal with large and noisy datasets.

Acknowledgements

DA is supported by KU Leuven CoE PFV/10/016 SymBioSys and IMEC HI^2 Data Science.

References

[BL13] BACHE K., LICHMAN M.: UCI machine learning repository, 2013. URL: http://archive.ics.uci.edu/ml. 2

[Cha05] CHAWLA N. V.: Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*. Springer, 2005, pp. 853–867. 1

[Dav02] DAVIDSON I.: Visualizing clustering results. In *Proceedings of the 2002 SIAM International Conference on Data Mining* (2002), SIAM, pp. 3–18. 1

[EKSX96] ESTER M., KRIEGEL H.-P., SANDER J., XU X.: A density-based algorithm for discovering clusters in large spatial databases with noise. AAAI Press, pp. 226–231. 1

[KJS96] KETCHEN JR D. J., SHOOK C. L.: The application of cluster analysis in strategic management research: an analysis and critique. *Strategic management journal* (1996), 441–458. 2

[Mac67] MACQUEEN J.: Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* (Berkeley, Calif., 1967), University of California Press, pp. 281–297. URL: http://projecteuclid.org/euclid.bsmsp/1200512992. 1

[SC04] SALVADOR S., CHAN P.: Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on* (2004), IEEE, pp. 576–584. 1