Cluster Correspondence Views for Enhanced Analysis of SOM Displays

Jürgen Bernard[*]
Interactive Graphics Systems Group
TU Darmstadt, Germany

Tatiana von Landesberger[†]
Interactive Graphics Systems Group
TU Darmstadt, Germany, and
Fraunhofer IGD, Darmstadt

Sebastian Bremm[‡]
Interactive Graphics Systems Group
TU Darmstadt, Germany

Tobias Schreck[§]
Interactive Graphics Systems Group
TU Darmstadt, Germany

Abstract

The Self-Organizing Map (SOM) algorithm [4] is a popular and widely used cluster algorithm. Its constraint to organize clusters on a grid structure makes it very amenable to visualization. On the other hand, the grid constraint may lead to reduced cluster accuracy and reliability, compared to other clustering methods not implementing this restriction. We propose a visual cluster analysis system that allows to validate the output of the SOM algorithm by comparison with alternative clustering methods. Specifically, visual mappings overlaying alternative clustering results onto the SOM are proposed. We apply our system on an example data set, and outline main analytical use cases.

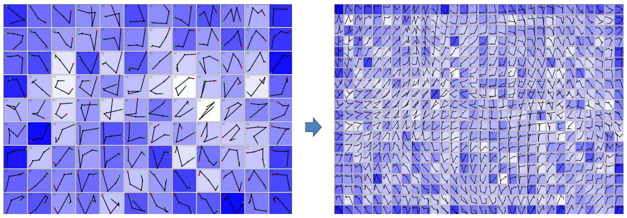
Index Terms: H.4 [Information Systems]: Information Systems Applications; I.3.6 [Computing Methodologies]: Methodology and Techniques

1 Introduction and Previous Work

Cluster analysis is a valuable data mining technique, supporting analysis of large data collections by abstraction to a limited number of data representatives (clusters). Visualization of cluster analysis results can help to better understand and communicate the output of the cluster algorithm at hand. While many clustering algorithms exist, the SOM method is very applicable to visualization, as it arranges the resulting clusters on a grid structure. However, this grid constraint at the same time may impact the algorithm's precision as compared to other methods that do not incorporate this constraint. The SOM method only provides implicit information about the number of clusters found in the data. Furthermore, the SOM method requires setting of a significant number of parameters. It may lead to increased uncertainty about the validity of the SOM results. For these reasons, it is also desirable to compare and validate the output of the SOM method against alternative clustering algorithms. Specifically, methods exist that provide reduced quantization error or allow better assessment of the actual number of clusters found. Integrating such methods with the SOM analysis in an appropriate way is expected to combine the advantages of each methods, while avoiding respective disadvantages.

In previous work, we introduced an interactive system that allowed to visually monitor and steer the SOM clustering process [5]. The system provides enhanced visual controls for the different parameters and is able to progressively visualize the resulting clustering result. However, the assessment of the quality of the SOM result is restricted to standard SOM-based diagrams such as distance or density maps [4]. We here present an extension to the available methods that allows to compare the SOM output with alternative clustering algorithms. It is based on mappings of clusters obtained by external clustering methods to the SOM cluster grid. We introduce our extensions and apply them to an example data set. Our approach is applicable to any data that can be processed by the SOM algorithm. In this poster, we use a data set consisting of *trajectories*, as considered in [5] (see Figure 1(a) for an illustration).

2 Interactive Cluster Correspondence Views



(a) SOM output

(b) Higher-resolution SOM output

Figure 1: SOM-based cluster analysis of a trajectory data set [5] showing the result grid of prototype trajectories with background color indicating cell densities (from light to dark color). Left: SOM clustering result. Right: Higher resolution SOM result using method of [2].

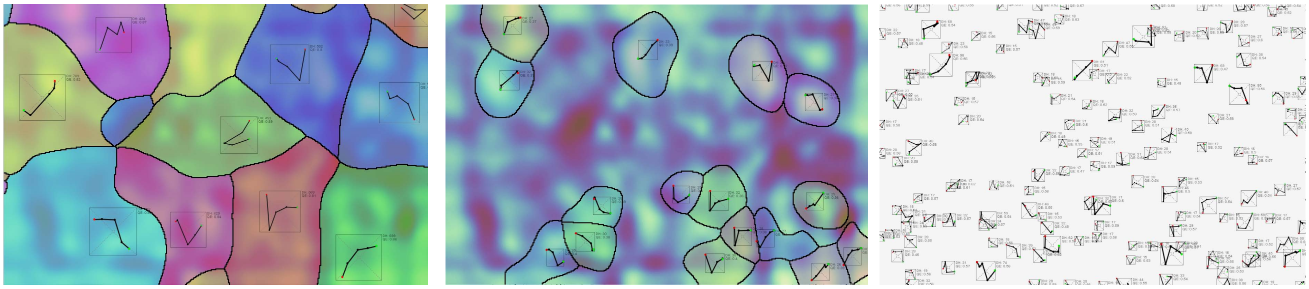
Our cluster correspondence view works on top of the SOM cluster grid as shown in Figure 1(b). It overlays SOM results by structures obtained by alternative clustering methods. The correspondence view can be chosen from a variety of proposed visualization methods. The overlay process is as follows:

1. Generation of alternative clusterings. Alternative clustering results are generated using selected clustering methods. We implemented the k-Means and DBScan methods [3] as alternative partitioning and density-based clustering algorithms, respectively. The methods are invoked by the setting of clustering parameters: the number of clusters k for k-Means, or density and distance thresholds for DBScan.

2. Mapping alternative clusters onto SOM reference grid. We map each obtained cluster onto the SOM reference grid by means of Spline-based continuous interpolation of the grid of SOM clusters. The technique was initially introduced to support mapping of large numbers of data points [2], but is equally suited for mapping alternative clusters for comparison purposes.

3. Visual representation of the alternative clusters. We propose several options to visually represent the relation of the

[*] e-mail: Juergen.Bernard@gris.informatik.tu-darmstadt.de
[†] e-mail: Tatiana.von_Landesberger@gris.informatik.tu-darmstadt.de
[‡] e-mail: Sebastian.Bremm@gris.informatik.tu-darmstadt.de
[§] e-mail: Tobias.Schreck@gris.informatik.tu-darmstadt.de



(a) k-Means correspondence view (b) DBScan correspondence view (c) Increased number of DBScan clusters

Figure 2: Overlaying alternative clustering results onto the SOM map for validation and refinement purposes. It shows two clustering algorithms with various correspondence visualizations. Left: K-means result overlay on SOM distance view. Color are assigned to clusters. Middle: DBScan result overlay on SOM distance view. Right: DBScan result. Icons show representative data elements for each cluster. Icon size indicates cluster size. Darker background shading means higher average distance to neighboring cells.

found alternative clusters to the underlying SOM grid. The representation comprises (a) a glyph for the cluster representative based on the underlying data, whose size represents the cluster size; and (b) a color-coding of corresponding areas on the SOM grid according to the cluster membership. Both representations are mapped onto the given SOM visualization as an additional layer of information.

These additional displays can be combined with the system's underlying SOM visualization capabilities. This is done by blending both visualizations, adding another layer of information to the base SOM display.

3 Application Examples

We demonstrate the capabilities of our system by application to a test dataset of 5000 trajectory elements. An initial SOM clustering of the data set is shown in Figure 1. It displays the SOM cluster grid where each cell shows a representative trajectory. The background color represents the density.

Figure 2(a) shows the mapping of a k-Means clustering ($k = 10$) onto the base SOM. The position of each cluster on the map is shown by a colored area (in the nearest neighbor sense). Each cluster is assigned a unique color. Each cluster is also represented by an icon showing a sample cluster trajectory. This view allows to assess the distribution of clusters over the map. As the k-Means is a partitioning scheme, the whole map is covered. Underlying the k-Means visualization layer is an original SOM view showing the average cell distance to its neighbors (light colors mean low distance). Owing to the opacity-based drawing, both cluster visualization layers can be jointly visually analyzed.

Figure 2(b) shows a mapping of a DBScan clustering result to the base SOM. As DBScan is a density-based scheme, therefore not all data items in the data set are necessarily also represented in the DBScan result. In particular, data from non-dense areas are considered as noise and are therefore ignored in the DBScan output. In this case, the 11 dominant clusters identified by the DBScan algorithm, cover only a part of the whole SOM map. This indicates that significant parts in the SOM map do not constitute relevant clusters, according to the particular parameterization. The view includes small trajectory icons showing the representative trajectory elements in the dataset. The icons size indicates the cluster size. We can use this display to easily assess the nature and distribution of DBScan clusters, in relation to the overall distribution of data elements. An overlay of this view with cell distance view (light colors mean small average distances to the neighbor cell) show a strong correspondence between density-based clustering result and the dense areas of SOM grid.

Figure 2(c) visualizes the result of a DBScan cluster analysis using a different parametrization from the previous example. In this case, a larger number of clusters was generated. The underlying SOM grid is used as a reference layout for positioning of the DBScan clusters. The scale of the cluster icons represents the size of the DBScan clusters. The display allows the analyst to concentrate on the DBScan clustering result, abstracting from the SOM result.

4 Conclusions and Future Work

We presented an extension of a SOM-based visual cluster system that allows to validate and assess the SOM cluster output by means of alternative clusterings that can be overlaid onto the SOM result grid. We demonstrated our method for two clustering algorithms: k-means and DBScan. The comparison of SOM output with k-Means clustering provides information of cluster membership with possibly better quantization accuracy as the SOM method. DBScan comparison identifies dense areas in the SOM cluster space. The integration of additional, alternative clusterings are expected to be useful for arriving at more reliable clustering results. Specifically, by our approach the user is able to validate her hypotheses about characteristics of clusters, by visually comparing the "opinions" of alternative clustering methods.

Future work includes incorporation of additional clustering algorithms, and design of cluster glyphs to encode addition statistical information about the found clusters and their interrelationships. We will also systematically explore approaches for visualization of quality aspects in visual cluster analysis. First steps toward this goal are taken in [1].

Acknowledgements

This work was partially supported by the German Research Foundation (DFG) within the project Visual Feature Space Analysis, as part of the Priority Program on Scalable Visual Analytics.

References

[1] J. Bernard, T. von Landesberger, S. Bremm, and T. Schreck. Multi-perspective quality views for improved visual cluster analysis with self-organizing maps. Submitted.

[2] J. Bernard, T. von Landesberger, S. Bremm, and T. Schreck. Micro-macro views for visual trajectory cluster analysis. In *Eurographics/IEEE Symposium on Visualization*, 2009. Poster.

[3] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2006.

[4] T. Kohonen. *Self-Organizing Maps*. Springer, 3rd edition, 2001.

[5] T. Schreck, J. Bernard, T. Tekušová, and J. Kohlhammer. Visual cluster analysis of trajectory data with interactive Kohonen maps. *Palgrave Macmillan Information Visualization*, 8:14–29, 2009.