Eurographics Conference on Visualization (EuroVis) (2015)
R. Maciejewski and F. Marton (Editors)

Quality Metrics Driven Approach to Visualize Multidimensional Data in Scatterplot Matrix

Michael Behrisch, Lin Shao, Juri Buchmüller and Tobias Schreck

University of Konstanz, Germany

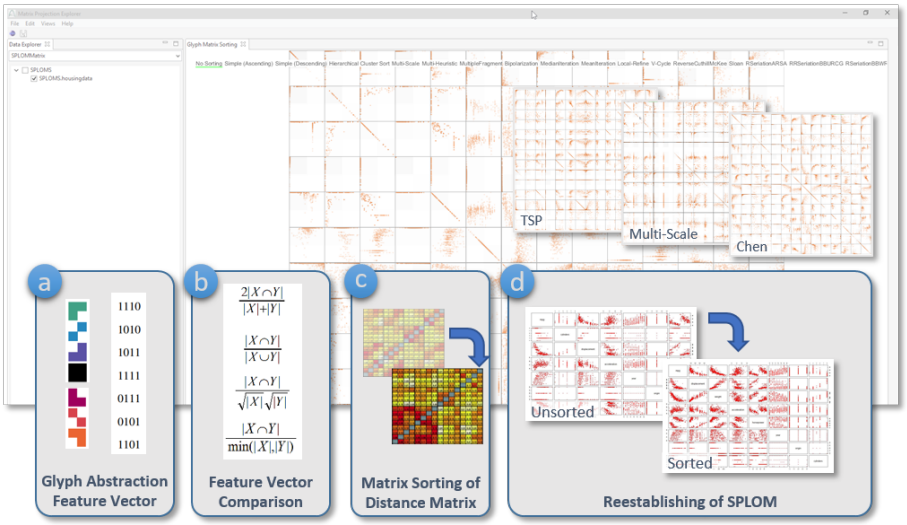

Figure 1: SPLOM Reordering Pipeline: Scatterplots are encoded by their visual motifs and encoded into a binary feature vector. A pair-wise comparison of all scatterplot motifs results in a distance matrix, which can be sorted with standard 2D numeric sorting algorithms (e.g., TSP- , Multi-Scale-, Chen ordering) to determine a visually coherent SPLOM ordering.

Abstract

Extracting meaningful information out of vast amounts of high-dimensional data is challenging. Prior research studies have been trying to solve these problems through either automatic data analysis or interactive visualization approaches. Our grand goal is to derive representative and generalizable quality metrics and to apply these to amplify interesting patterns as well as to mute the uninteresting noise for multidimensional visualizations. In this poster, we investigate a quality metrics-driven approach to achieve our goal for scatterplot matrices (SPLOMs). We rearrange SPLOMs by sorting scatterplots based on their locally significant visual motifs. Using our approach, we enable scatterplot matrices to reveal groups of visual patterns appearing adjacent to each other, helping analysts to gain a clear overview and to delve into specific areas of interest more easily.

Categories and Subject Descriptors (according to ACM CCS): Information Systems [H.5.0]: Information Interfaces and Presentation—General;

1. Introduction

Extracting meaningful information out of vast amounts of high-dimensional data is a challenging task. General exploration- and retrieval tasks, such as finding relevant dimensions, selecting meaningful projections, or investigating outliers, are significantly more challenging in high-dimensional data analysis. Multi-dimensional data visualization also carries its own set of challenges like, above all, the

limited capability of any technique to scale to more than a couple of data dimensions. Prior research studies developed many visualization techniques to achieve the goal, such as parallel coordinates, scatterplots, and glyphs. However, mere visualization of all variables may introduce clutter and blurs interesting patterns in visualizations.
Researchers have been trying to solve these problems through either automatic data analysis or interactive visualization approaches. We propose a mixed approach, where the system – based on *quality metrics* – automatically searches through a large number of potentially interesting views, and the user interactively steers the process and explores the output through visualizations. Our grand goal is to derive quality metrics, which amplify interesting patterns and mute the uninteresting noise for multidimensional visualizations.

In this poster, we investigate a quality metrics driven approach for scatterplot matrices (SPLOMs). In past decades, many dimension management techniques have been proposed to organize layouts automatically or interactively. Ankerst et al. [ABK98] proposed to place similar dimensions close together based upon similarity metrics. Dimension reordering can also be used to maximize the clarity of visual patterns in scatterplot matrices by reducing unnecessary clutter [PWR04, YPWR03]. In relation to visual patterns, Dang and Wilkinson [DW14] used Scagnostics (Scatterplot Diagnostics) to reveal hidden patterns in large collections of scatterplots. Visual cluster verification was empirically studied in [SMT13] to determine the impact of dimension reduction techniques and different scatterplot encodings (2D, 3D and SPLOMs). Interactive approaches, such as in [EDF08], propose to navigate and rearrange multidimensional data based upon iteratively built queries in scatterplot matrices. Despite lacking in the definition of the quality measurements, the quality-aware sorting framework for scatterplot matrices was also suggested [AEL*09]. Inspired by aforementioned techniques, this paper proposes quality metrics and an initial framework for quality metrics driven sorting for scatterplot matrices.

2. Quality Metrics from Visual Space

In comparison to the earlier approaches, we intend to use quality metrics derived from the visual space rather than the data space. In a SPLOM, the distribution of general patterns, called *scatterplot motifs*, are more interesting than the point distribution within one scatterplot cell. Hence, the effectiveness of a SPLOM, like many other matrix visualizations, is affected by its ordering. Thus, finding a good SPLOM ordering helps to reveal motif patterns and their distributions regardless of the dimensions under consideration.

Our approach aims to improve the visual coherence in SPLOMs by reordering the matrix, such that adjacent cells appear visually similar to each other and motifs groups form structural patterns. To demonstrate our approach, we created several SPLOMs of the UCI housing [Lic13] dataset with different orderings, as shown in Figure 1. It can be seen that the different matrix sorting algorithms promote different patterns (e.g., Multi-Scale groups line motifs, while TSP and Chen group similar patterns in adjacent locations).

3. The Pipeline of Our Approach

Our approach to find a visually coherent SPLOM ordering is as following: 1) we calculate visual similarity between scatterplots, and 2) we compare all scaterplots using the similarity score, which determines the final SPLOM ordering. Our approach for the ordering process is depicted in Figure 1.

Abstraction-Based Scatterplot Feature Descriptor Inspired by the work of Yates et al. [YWS*14], we abstract the scatterplots by their contained scatterplot motifs. In case of a 2×2 grid, 16 unique motifs can be derived and encoded in a binary vector form. In this vector, a 1 represents a scatterplot segment with a point density above a user selected threshold. Using the coding scheme in [YWS*14], we form a space-filling z-curve to traverse the scatterplot segments. Users may adjust grid sizes to steer the ordering process in the feature descriptor approach.

Feature Descriptor Comparison The binary feature vector, representing a scatterplot motif, allows comparing visual appearances using overlap comparison approaches. As Figure 1 (b) depicts, we can calculate similarity scores based on the Dice-, Jaccard-, Cosine-, and Overlap coefficients.

Distance Matrix Sorting As Figure 1 (c) illustrates, a pairwise calculation of the visual distances results in a distance matrix. Every cell in this symmetric matrix corresponds to the visual similarity score of the "pivot" scatterplot to another "comparison" scatterplot. We can apply a wide range of matrix sorting algorithms to reorder the numeric distance matrix. Currently, we are experimenting with the *R package Seriation* to obtain an implementation of the matrix sorting algorithms (see also [HHB08]).

Reestablishing of the SPLOM The sorted distance matrix can be directly translated back into its ordered SPLOM correspondence or into a sorted Glyph Matrix. Therefore, we retrieve the distance matrix ordering vector and apply it to the SPLOM rows and columns. Hence, the scatterplot with the highest or lowest–depending on the matrix sorting algorithm–visual similarity to the rest of scatterplots are placed in the top-left corner of the SPLOM, as shown in Figure 1 (d). Other scatterplots are subsequently arranged with respect to their distance values.

4. Conclusion

Our main goal is to use quality metrics to improve scatterplot matrices. In the scope of this paper, we shared our approach and pipeline to rearrange scatterplot matrices using scatterplot motifs as described in [YWS*14]. Using the idea, we expect scatterplot matrices to reveal groups of visual patterns appearing adjacent to each other, which helps analysts to gain a clear overview and to delve into specific areas of interest more easily. Our ongoing investigation aims to test and refine the feature vector for scatterplot motifs depending upon data sizes and the number of dimensions.

References

[ABK98] ANKERST M., BERCHTOLD S., KEIM D.: Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In *IEEE Symposium on Information Visualization, 1998. Proceedings* (Oct. 1998), pp. 52–60, 153. 2

[AEL*09] ALBUQUERQUE G., EISEMANN M., LEHMANN D. J., THEISEL H., MAGNOR M. A.: Quality-based visualization matrices. In *Proceedings of the Vision, Modeling and Visualization (VMV)* (2009), pp. 341—350. 2

[DW14] DANG T. N., WILKINSON L.: Transforming scagnostics to reveal hidden features. *Visualization and Computer Graphics, IEEE Transactions on 20*, 12 (Dec 2014), 1624–1632. 2

[EDF08] ELMQVIST N., DRAGICEVIC P., FEKETE J.: Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE Transactions on Visualization and Computer Graphics 14*, 6 (Nov. 2008), 1539–1148. 2

[HHB08] HAHSLER M., HORNIK K., BUCHTA C.: Getting things in order: An introduction to the r package seriation. *Journal of Statistical Software 25*, 3 (March 2008), 1–34. 2

[Lic13] LICHMAN M.: UCI machine learning repository, 2013. URL: http://archive.ics.uci.edu/ml. 2

[PWR04] PENG W., WARD M., RUNDENSTEINER E.: Clutter reduction in multi-dimensional data visualization using dimension reordering. In *IEEE Symposium on Information Visualization, 2004. INFOVIS 2004* (2004), pp. 89–96. 2

[SMT13] SEDLMAIR M., MUNZNER T., TORY M.: Empirical guidance on scatterplot and dimension reduction technique choices. *Visualization and Computer Graphics, IEEE Transactions on 19*, 12 (Dec 2013), 2634–2643. 2

[YPWR03] YANG J., PENG W., WARD M., RUNDENSTEINER E.: Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In *IEEE Symposium on Information Visualization, 2003. INFOVIS 2003* (Oct. 2003), pp. 105–112. 2

[YWS*14] YATES A., WEBB A., SHARPNACK M., CHAMBERLIN H., HUANG K., MACHIRAJU R.: Visualizing multidimensional data with glyph sploms. *Computer Graphics Forum 33*, 3 (2014), 301–310. 2