Quality Metrics Driven Approach to Visualize Multidimensional Data in Scatterplot Matrix

Michael Behrisch* Lin Shao† Bum Chul Kwon‡ Tobias Schreck§ Ivan Sipiran¶ Daniel Keim‖

University of Konstanz

ABSTRACT

Extracting meaningful information out of vast amounts of high-dimensional data is very difficult. 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 the representative and generalizable quality metrics and to apply the metrics to amplify interesting patterns as well as to mute the uninteresting noise for multidimensional visualizations. In this particular poster, we investigate quality metrics driven approach to achieve the goal for scatterplot matrix (SPLOM). Our main approach is to rearrange scatterplot matrices by sorting scatterplots based upon their patterns especially locally significant ones, called scatterplot motifs. Using the approach, 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.

Index Terms: H.5.0 [Information Systems]: Information Interfaces and Presentation—General;

1 INTRODUCTION

Extracting meaningful information out of vast amounts of high-dimensional data is very difficult. The curse of dimensionality is a popular way of stigmatizing the whole set of issues encountered in high-dimensional data analysis: finding relevant projections, selecting meaningful dimensions, removing noise, and being only a few of them. 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 handful of data dimensions. Prior research studies developed many visualization techniques to achieve the goal, such as parallel coordinates, scatterplots, glyphs, and dimension stacking. However, mere visualization of all variables in such ways may produce too much clutter, which often blurs interesting patterns in visualizations.

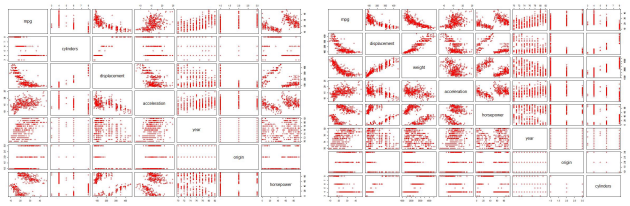
Researchers have been trying to solve these problems through either automatic data analysis or interactive visualization approaches. What is needed is a mixed approach, where the machine-based on quality metrics-automatically searches through a large number of potentially interesting transformations and projections, and the user interactively steers the process and explores the output through visualizations. Our grand goal is to derive the representative and generalizable quality metrics and to apply the metrics to amplify interesting patterns as well as to mute the uninteresting noise for multidimensional visualizations.

In this particular poster, we investigate quality metrics driven approach for scatterplot matrix (SPLOM). In past decades, many dimension management techniques have been proposed to organize layouts automatically or interactively. Ankerst et al. [2] proposed to place similar dimensions close together based upon similarity metrics. In addition, a hierarchical dimension ordering, spacing, and filtering approach automatically arranges dimensions based upon dimension similarities and allows users to interactively explore them [6]. Dimension reordering can also be used to maximize the clarity of visual patterns in scatterplot matrix by reducing unnecessary clutter [5]. Elmqvist et al., [3] proposes an interactive approach to navigate and rearrange multidimensional data based upon iteratively built queries in scatterplot matrix. Despite lack of definition of the quality measurements, the quality-aware sorting framework for scatterplot matrix was also suggested [1]. Inspired by aforementioned techniques, this paper proposes quality metrics and an initial framework for quality metrics driven sorting for scatterplot matrix.

2 BASIC IDEA

In comparison to the earlier approaches, we intend to use quality metrics derived from the visual space, in stead of the data space. In a SPLOM, the distribution of general patterns, *scatterplot motifs*, are of interest rather than the point distribution within one scatterplot cell. Hence, the effectiveness of a SPLOM–alike any other matrix visualization–is directly related to its ordering. Accordingly, finding a good SPLOM ordering helps to reveal motif patterns, or groups thereof, and their distributions regardless of the dimensions under consideration.

Our approach targets to improve the visual coherency in SPLOMs by reordering the matrix, such that adjacent cells appear visually similar to each other. In such an ordered SPLOM, visually similar motifs consitute structural patterns. As shown in Figure 1, the scatterplots containing similar patterns appear in adjacent locations: horizontal lines at the bottom, vertical lines on the right side, and clearly "linear" correlations on the upper left side.



(a) Randomly positioned dimensions (b) Ordered dimensions

Figure 1: Scatterplot matrices visualization of Cars dataset. (a) dimensions are randomly positioned. (b) the SLOM cells are sorted allowing to perceive dataset patterns.

*e-mail:michael.behrisch@uni-konstanz.de
†e-mail: lin.shao@uni-konstanz.de
‡e-mail:bumchul.kwon@uni-konstanz.de
§e-mail:tobias.schreck@uni-konstanz.de
¶e-mail:sipiran@dbvis.inf.uni-konstanz.de
‖e-mail:daniel.keim@uni-konstanz.de

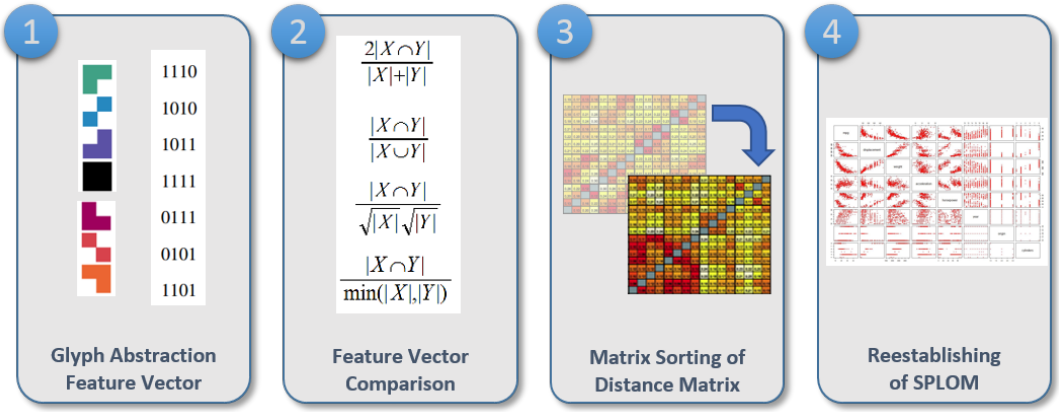


Figure 2: SPLOM Reordering Pipeline: Scatterplots are encoded by their visual motifs, such as presented in [7]. The motifs can be encoded into a binary feature vector. A pair-wise comparison of all scatterplot motifs results in a distance matrix. This distance matrix can be sorted with standard 2D numeric sorting algorithms, such as TSP Ordering, Optimal-Leaf Ordering, or the Sloan sorting algorithm. In the final step, the distance matrix's ordering can be applied to the initial SPLOM to achieve a visually coherent SPLOM ordering.

3 Approach / Pipeline

Our main 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. We derive a pipeline approach for the ordering process, as depicted in Figure 2.

Abstraction-Based Scatterplot Feature Descriptor Inspired by the work of Yates et al. [7] we abstract the scatterplots by their contained scatterplot motifs. In case of a 4×4 grid, 15 unique motifs can be derived that can be described in a binary vector form. Every index in this vector of length four contains 1 if and only if its scatterplot segment has a point density above a user selected threshold. Using the coding scheme in [7], we form a space-filling z-curve, starting from the top left, to traverse the scatterplot segments. For example, the upper (green) scatterplot motif in Figure 2-(a) is the visual depiction of the feature vector $< 1110 >$. Users may adjust grid sizes to steer the ordering process in the feature descriptor approach.

Feature Descriptor Comparison Since the scatterplot is described by a binary feature vector representing its scatterplot motif, we can compare the visual appearances using standard overlap comparison approaches. Several possibilities can be chosen by users. As Figure 2-(a) depicts, we can calculate similarity scores based on the Dice-, Jaccard-, Cosine-, Overlap coefficients.

Distance Matrix Sorting As Figure 2-(c) illustrates, a pair-wise calculation of the visual distances between all scatterplots of the SPLOM results in a distance matrix. Here, every cell in this symmetric matrix corresponds to the visual similarity score of the "pivot" scatterplot to another "comparison" scatterplot. Since the distance matrix contains only numeric entries, we can apply a wide range of matrix sorting algorithms to reorder the distance matrix. Currently, we are experimenting with the *R package Seriation* to obtain an implementation of the matrix sorting algorithms (see also [4]).

Reestablishing of the SPLOM The sorted distance matrix can be directly translated back into its ordered SPLOM correspondence or into a sorted Glyph Matrix [7]. Therefore, we retrieve the vector that describes the new distance matrix order from the matrix sorting algorithm and apply the vector directly to the SPLOM columns and rows. Hence, the scatterplot with the highest or the lowest–depending on the matrix sorting algorithm–visual similarity to the rest of scatterplots are placed in in the top-left corner of the SPLOM. Other scatterplots are subsequently arranged with respect to their distance values.

4 Conclusion

Our main goal is to use quality metrics to visualize multidimensional data in scatterplot matrices. In the scope of this paper, we shared our motivation and approaches to rearrange scatterplot matrices using the visual quality, scatterplot motifs. Using the approach, 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

[1] G. Albuquerque, M. Eisemann, D. J. Lehmann, H. Theisel, and M. A. Magnor. Quality-based visualization matrices. In *Proceedings of the Vision, Modeling and Visualization (VMV)*, pages 341—350, 2009.

[2] M. Ankerst, S. Berchtold, and D. Keim. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In *IEEE Symposium on Information Visualization, 1998. Proceedings*, pages 52–60, 153, Oct. 1998.

[3] N. Elmqvist, P. Dragicevic, and J. Fekete. Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1539–1148, Nov. 2008.

[4] M. Hahsler, K. Hornik, and C. Buchta. Getting things in order: An introduction to the r package seriation. *Journal of Statistical Software*, 25(3):1–34, March 2008.

[5] W. Peng, M. Ward, and E. Rundensteiner. Clutter reduction in multidimensional data visualization using dimension reordering. In *IEEE Symposium on Information Visualization, 2004. INFOVIS 2004*, pages 89–96, 2004.

[6] J. Yang, W. Peng, M. Ward, and E. Rundensteiner. Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In *IEEE Symposium on Information Visualization, 2003. INFOVIS 2003*, pages 105–112, Oct. 2003.

[7] A. Yates, A. Webb, M. Sharpnack, H. Chamberlin, K. Huang, and R. Machiraju. Visualizing multidimensional data with glyph sploms. *Computer Graphics Forum*, 33(3):301–310, 2014.