Identifying Locally Interesting Motifs for Exploration of Scatter Plot Matrices

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

University of Konstanz

ABSTRACT

Scatter plots are effective diagrams to visualize distributions, clusters and correlations in two-dimensional data space. For high-dimensional data, scatter plot matrices can be formed to show all two-dimensional combinations of dimensions. Several previous approaches for exploration of large scatter plot spaces have focused on ranking and sorting scatter plot matrices based on global patterns. However, often local patterns are of interest for scatter plot exploration. We present a preliminary idea to explore the scatter plot space by identifying significant local patterns (also called motifs in this work). Based on certain clustering algorithms and image-based descriptors, we identify and group a set of similar local candidate motifs in a large scatter plot space.

Keywords: High-dimensional data analysis, similarity measure, pattern recognition, scatter plots.

1 INTRODUCTION

Nowadays, vast amount of data are created in many application domains and contain potentially interesting information, which need to be explored and analyzed. However, due to the fact that data grows rapidly, the problem of effective and efficient access to large multivariate and high-dimensional data arises. While in the past, the storage of large amounts of data was the primary problem, today the challenge focuses e.g., on detecting interesting patterns and insights in large data repositories. *Scatter plot* visualizations are one of the most widely used and well-understood visual representation for bivariate data. They can be applied also for high-dimensional data via dimensionality reduction or the scatter plot matrix (SPLOM) representation [4]. However, since the task to find interesting scatter plots in high-dimensional data is referring to an n-dimensional problem, the exploration and perception of single scatter plots may suffer for large scatter plot matrices.

Manually searching through a large space of possible views or data sections is expensive and may become infeasible for large or high-dimensional data sets. Recent work in Visual Analytics has focused on computing quality metrics [1] which can be used to filter and rank large data spaces to present the user a good starting point for exploration. Specifically, several previous works [2] have focused on interestingness measures based on *global* properties of scatter plots for ranking and ordering. In this work we contribute a novel exploration tool to discover high-dimensional dependencies and to find interesting dimension combinations based on local motifs. The problem of finding local interestingness measures for scatter plots is closely related to interest point detection. Visually discriminating motifs are considered the basis for extraction of interest measures, since they can be quickly recognized by the human and furthermore, represent relationships between involved dimensions. By means of this approach large scatter plot matrices can be explored or filtered for interesting local point distributions.

*e-mail: lin.shao@uni-konstanz.de

†e-mail:michael.behrisch@uni-konstanz.de

‡e-mail:tobias.schreck@uni-konstanz.de

§e-mail:sipiran@dbvis.inf.uni-konstanz.de

¶e-mail:bumchul.kwon@uni-konstanz.de

‖e-mail:daniel.keim@uni-konstanz.de

2 BASIC IDEA

The basic idea of our visual exploration tool is to help the users to investigate similar local motifs in an unsorted SPLOM rather than to reorder SPLOMs. Therefore, we apply a dictionary-based approach to identify similar motif candidates from scatter plot space. Our approach is based on the following pipeline:

1. First, a local scatter plot segmentation step is applied. It extracts dense point regions from global scatter plots.

2. Then, image-based descriptors are used to extract the visual features of all identified motifs.

3. Based on the feature similarity, a dictionary of motifs is computed to represent the entire scatter plot space.

Since we are looking for clearly separated motifs, we decided to use a DBSCAN algorithm to extract locally dense point distributions as local motifs. This is one option to extract the local motifs, of course, there are plenty of other possibilities conceivable to achieve this. After we identified all local segments, we compute image-based feature vectors which describe the density distribution and the general shape by an edge detection approach [3]. By means of the image-based feature vectors, we may cluster the motifs regardless of position or axes scales. Finally, a motif-based dictionary is formed by a k-means clustering on all local feature vectors.

3 APPROACH / VISUALIZATION

An essential element of this approach is to find an appropriate parameter setting for k which defines the number of dictionary entries. Since the number of dictionary entries highly depends on the given data set, it is necessary to vary this parameter. Therefore, we have developed a visual exploration tool to discover a good parameter for k to build up a dictionary and to analyze the similar identified candidate motifs. The exploration tool involves a global overview in the form of a SPLOM and a detailed tabular view of all clustered motifs. Furthermore, it allows the user to experiment with different image-based descriptors which also influence the result.

In Figure 1, we demonstrate an exemplified interestingness measure based on a motif dictionary with 39 entries computed by means of an *Edge Histogram Descriptor* [3]. The used data set was provided by the *Movebank*[1] repository and contains animal tracking data including a range of transmitters' reports of geolocations and electronic measurements of pressure, temperature or wind properties. The data set consists of 5000 data records and 21 dimensions. From this input data set, our segmentation approach returned 2012 local motifs of a total of 210 scatter plots.

[1] http://www.movebank.org

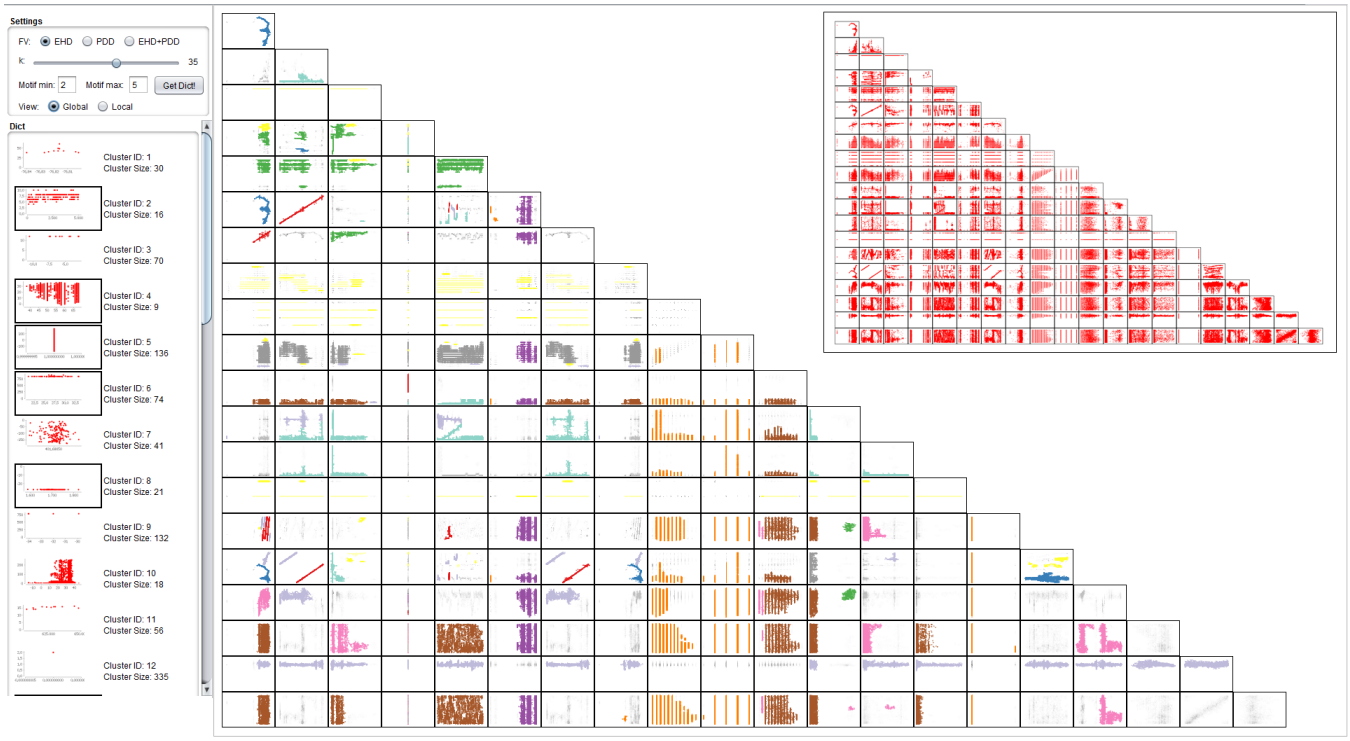

Figure 1: An Overview of our visual motif identification tool to find similar local motifs, as well as a good parameter setting for the motif-based dictionary.

Once a good parameter setting has been found, the user can explore the SPLOM for interesting motifs. By clicking on a dictionary entry all cluster members will be highlighted in the linked SPLOM. Conversely, it is possible to highlight all related motifs by clicking on a given motif in the scatter plot space. Moreover, to distinguish a multiple selection, we use different color codings for each dictionary and motif respectively. By means of this visual indication the user can quickly recognize the dictionary quality, motif distribution in the scatter plot space and whether the chosen settings are suitable for a given data set or not. A further benefit of this overview is that the user can estimate if the number of clusters is rather too high or too low for the given scatter plot space.

Figure 1 illustrates one example of an appropriate dictionary setting, once with (center) and once without (upper right corner) highlighting local motifs. To demonstrate the effectiveness, we highlight 11 of 35 significant dictionary entries and may see the diverse distributions of the motifs by different color mappings. Moreover, it reveals that certain motifs just appear in specific rows or columns (e.g., light grey motifs in the penultimate row or the orange motifs in column nine and ten). By considering all local motif of particular rows or columns we are able to identify dimension dependencies and may filter identical dimension combinations for a more compact motif exploration. In some cases, when the number of clusters k is chosen too small, dissimilar motifs will be aggregated to one dictionary entry and the corresponding cluster representative cannot represent the entry well. Hence, we provide an additional local motif view to observe all clustered motifs according to the dictionary entries.

4 Conclusion and Future Work

We presented a novel visual exploration tool to identify user-adaptive interesting motifs over different dimension combinations in a SPLOM. It is based on a motif-dictionary approach which groups and highlights all similar motifs in the SPLOM.

Furthermore, this approach can be extended in different directions. An automatic ranking based on the dictionary approach could be applied to reorder SPLOMs or to compute an overall interesting (global) score based on the frequency of local motifs. Heterogeneous data sets may be analyzed by user-adaptive motifs. Textual data could be interesting to this end. In addition, we plan to further investigate for extraction and analysis of local motifs. Enhanced high-dimensional visualization techniques could help to visualize local motifs in combination with other data types.

Acknowledgements

This work was partially funded by the Juniorprofessor Program of the Landesstiftung Baden-Württemberg within the research project *Visual Search and Analysis Methods for Time-Oriented Annotated Data.*

References

[1] G. Albuquerque, M. Eisemann, D. J. Lehmann, H. Theisel, and M. Magnor. Quality-Based Visualization Matrices. *In Proceedings of the Vision, Modeling and Visualization*, pages 341–350, 2009.

[2] W. Peng, M. O. Ward, and E. A. Rundensteiner. Clutter reduction in multi-dimensional data visualization using dimension reordering. *Proceedings - IEEE Symposium on Information Visualization, INFO VIS*, pages 89–96, 2004.

[3] L. Shao, M. Behrisch, T. Schreck, T. von Landesberger, M. Scherer, S. Bremm, and D. A. Keim. Guided Sketching for Visual Search and Exploration in Large Scatter Plot Spaces. In *Proc. EuroVA International Workshop on Visual Analytics*, 2014.

[4] M. Ward, G. Grinstein, and D. Keim. *Interactive Data Visualization: Foundations, Techniques, and Applications*. A. K. Peters, Ltd., Natick, MA, USA, 2010.