Journal of Visual Languages and Computing 36 (2016) 1–12

Contents lists available at ScienceDirect

Journal of Visual Languages and Computing

journal homepage: www.elsevier.com/locate/jvlc

Guiding the exploration of scatter plot data using motif-based interest measures

Lin Shao [a,*], Timo Schleicher [b], Michael Behrisch [b], Tobias Schreck [a], Ivan Sipiran [c], Daniel A. Keim [b]

[a] Graz University of Technology, Graz, Austria
[b] University of Konstanz, Konstanz, Germany
[c] Pontificia Universidad Católica del Perú PUCP, Lima, Peru

ARTICLE INFO

Article history:
Received 28 February 2016
Accepted 6 July 2016
Available online 17 July 2016

MSC:
62-07
68U05

Keywords:
Scatter plot
Local patterns
Motifs
Visual dictionary

ABSTRACT

Finding interesting patterns in large scatter plot spaces is a challenging problem and becomes even more difficult with increasing number of dimensions. Previous approaches for exploring large scatter plot spaces like e.g., the well-known Scagnostics approach, mainly focus on ranking scatter plots based on their *global* properties. However, often *local* patterns contribute significantly to the interestingness of a scatter plot. We are proposing a novel approach for the automatic determination of interesting views in scatter plot spaces based on analysis of local scatter plot segments. Specifically, we automatically classify similar local scatter plot segments, which we call scatter plot *motifs*. Inspired by the well-known $tf \times idf$-approach from information retrieval, we compute local and global quality measures based on frequency properties of the local motifs. We show how we can use these to filter, rank and compare scatter plots and their incorporated motifs. We demonstrate the usefulness of our approach with synthetic and real-world data sets and showcase our data exploration tools that visualize the distribution of local scatter plot motifs in relation to a large overall scatter plot space.

Contents

* Corresponding author.
E-mail addresses: l.shao@cgv.tugraz.at (L. Shao), timo.schleicher.edu@gmail.com (T. Schleicher), michael.behrisch@uni-konstanz.de (M. Behrisch), tobias.schreck@cgv.tugraz.at (T. Schreck), iasipiranm@gmail.com (I. Sipiran), keim@uni-konstanz.de (D.A. Keim).

http://dx.doi.org/10.1016/j.jvlc.2016.07.003
1045-926X/

1. Introduction

Nowadays, vast amounts of data are rapidly created in many application domains and thus the problem of effective and efficient access to large multivariate and high-dimensional data arises. While in the past, the storage capacity was the primary problem, today the challenges comprise tasks like detecting interesting patterns or correlations in large data sets. One solution is to apply suitable visualization techniques and search for hidden information within the data. *Scatter plot* visualizations are one of the most widely used and well-understood visual representations for bivariate data. They can also be applied for high-dimensional data via dimensionality reduction or the scatter plot matrix representation [1]. However, perceiving and finding interesting scatter plots in large scatter plot collections constitutes a severe challenge, especially when working with scatter plot matrices.

Manually searching through large amounts of data views is exhaustive and may become infeasible for high-dimensional data sets. Recent work in Visual Analytics has focused on computing interestingness measures, which can be used to filter and rank large data spaces to present the user a good starting point for exploration. Specifically, several previous approaches, such as [2–4], have focused on interestingness measures based on *global* properties of scatter plots for ranking and filtering. However, global interesting scores do not consider the impact of local patterns, which add to the overall interestingness of a scatter plot. Often, it is a combination of several different local scatter plot patterns which by their composition constitute interesting data views.

Here, we present a novel approach to discover interesting scatter plot views, which opposed to current quality metrics focuses on scatter plot interestingness derived from *local* data properties. We adapt a minimum spanning tree-based clustering technique for a non-parametric segmentation of scatter plots as data preprocessing. Next, we apply ideas from the image analysis domain to scatter plots. Specifically, we extract visual features as the basis for clustering local scatter plot segments into groups of similar patterns, called motifs. Consequently, we are able to compute an interestingness measure for scatter plots in terms of the distribution of occurring motifs. Our idea here is that visually discriminating motifs are considered of interest, since they can be quickly recognized by the human. We apply a Bag-of-Visual-Words [5] concept for scatter plots and transfer the idea of $tf \times idf$-weighting to this domain. Thus, we can derive the interestingness of a local scatter plot motif based on its occurrence within a given scatter plot and in relation to the occurrence in other scatter plots of a scatter plot space. We make use of these local motif-based measurements to rank and filter large scatter plot spaces.

We claim the following technical contributions:

- We adapt the minimum spanning tree-based clustering technique for a non-parametric segmentation of scatter plot diagrams.
- We introduce a motif-based dictionary to assess the interestingness of local scatter plot patterns.
- We define a global interestingness score based on the occurrence and similarity of local motifs.

This work is a revised version of an earlier conference article [6]. In this paper, we contribute several extensions to our initial approach as follows. First, we extended the related work by discussion of recent papers on visual abstractions of scatter plot spaces and generation of projection views from high-dimensional input data. Furthermore, we also introduce technical extensions which improve the exploration of local scatter plot patterns. Specifically, the distribution of local patterns from the local pattern dictionary can be visually explored by a user-configurable Star Coordinate view including detail-on-demand. Also, we introduce a hybrid design for embedding scatter plot motif views within a Parallel Coordinate display, allowing to relate local scatter plot patterns with further data dimensions. The made extensions provide additional contributions and improve the usefulness and analytical power of the proposed approach.

The remainder of this paper is structured as follows: in Section 2, we discuss related work and show commonalities and highlight differences. Section 3 gives an overview of our general idea to use local motif analysis for computing local and global interestingness measures. In Section 4, we present technical details of our approach. Section 5 gives an overview of our visual exploration tools to identify and analyze local motifs. Next, in Section 6, we apply our implementation to different data sets and showcase a local motif-driven exploration. Our approach is only a first step to scatter plot analysis based on local patterns, and we discuss limitations and a range of possible extensions in Section 7. Finally, Section 8 concludes the paper.

2. Related work

Several works support the exploration of large scatter plot data sets by means of ranking, filtering and searching functionalities. We next review a selection of works in the context of our approach.

2.1. Visualization of scatter plot patterns

Visualizations of scatter plots need to have an appropriate aspect ratio and scale to reveal correlations, patterns, trends and clusters. This is challenging since the identification of patterns in scatter plots, and the notion of interestingness, are subjective in nature and depend on scale and proportions. Most existing aspect ratio optimization methods rely on properties of line segments displayed in a plot. In [7], it is suggested to use segments of a virtual polyline that connects all existing data points of a scatter plot, or the segments of a regression line through the plot. Talbot et al. [8] showed that this approach is suitable for data containing trends, but may be less appropriate for data which do not have this kind of functional relationship. Hence, they proposed a method based on contour lines resulting from a kernel density estimation, which is able to deal with pairs of variables without functional relationship. In a recent approach, Fink et al. [9] presented a scatter plot aspect ratio calculation that is based on the Delaunay triangulation of the data points. The authors claimed that the aspect ratio is appropriate if the edges of the Delaunay triangulation have certain geometric properties. In [10] a visual separation measure based on extended minimum spanning tree was presented to derive local patterns in projection mappings.