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, Perú*

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.

Keywords: Scatter plot, local patterns, motifs, visual dictionary

Email addresses: `l.shao@cgv.tugraz.at` (Lin Shao), `timo.schleicher.edu@gmail.com` (Timo Schleicher), `michael.behrisch@uni-konstanz.de` (Michael Behrisch), `tobias.schreck@cgv.tugraz.at` (Tobias Schreck), `iasipiranm@gmail.com` (Ivan Sipiran), `keim@uni-konstanz.de` (Daniel A. Keim)

2010 MSC: 62-07, 68U05

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, 3, 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 constitutes 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 discriminatory 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 its occurrence among and within the scatter plot corpus. 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.

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 implementation details. 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 [6], 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. [7] 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. [8] present a scatter plot aspect ratio calculation that is based on the Delaunay triangulation of the data points. The authors claim that the aspect ratio is appropriate if the edges of the Delaunay triangulation have certain geometric properties. In [9] a visual separation measure based on extended minimum spanning tree was presented to derive local patterns in projection mappings.

Another well-known problem of scatter plots is the degree of overlapping and overdrawing data points, which makes the identification of subgroups more difficult. In [10], an abstraction approach was introduced to group dense data points and to reveal relationships between subgroups by using smooth contour lines in combination with different color codings. Another recent work on visual abstraction has been presented by Chen at al. [11], where a multi-class sampling technique is presented that reduces the overdraw and preserves the point distributions for quantitative analysis. More generally, a study on perceptional factors, which links scatter plot properties with perceived interestingness and interpretability is given in [12].

2.2. Feature-Based Analysis of Scatter Plots

Automatic identification of interesting candidates within large sets of scatter plots has recently been an active field of research. The Scagnostics method [2] is a well-known feature-based approach, which proposes a set of graph-based measures for scatter plots, to describe the data properties. While the Scagnostics method does not require classified data, consistency measures [13] can further improve the identification of informative scatter plots for the case that class labels are available. In [14], a multi-step analysis of large scatter plot matrix spaces was introduced. The approach is based on visual quality measures, matrix reordering, and visual abstraction, and supports navigation and analysis in large scatter plot data.

Often, different scatter plot views need to be compared. In [15], two-dimensional color-coding was applied to compare sets of scatter plots for topological relationships. Other works supported the comparison of sets of scatter plots by automatic and interactive approaches. Albuquerque et al. [16] introduced an importance-aware sorting algorithm to find good projections in scatter plot matrices. A recently tackled problem is the identification of interesting subspaces in high-dimensional data, using scatter plots of projected subspaces. In [17], a sampling approach was shown that identifies interesting subspace projections for high-dimensional data sets. In [18], a visual approach for the identification of interesting subspaces was proposed. It relies on a clustering-based subspace search method to compute the interestingness score from density and class-separation measures.

2.3. Navigation in Scatter Plot Space

The effectiveness of analyzing large scatter plot data also depends on appropriate navigation facilities. Animated navigation and extrusion-based transitions between views was proposed in [19] to navigate in scatter plot matrix spaces. Scherer et al. [20] introduced a search and navigation interface that is based on the scatter plots global regression features. In [21], we introduced a supervised sketching system to search for interesting patterns in a large scatter plot space. We used image-based features to compare the similarity of user sketches and data patterns, and provide clustering methods to analyze associated dimensions. Another possibility to explore and navigate through scatter plot spaces is the usage of projection visualizations in connection with extracted features. For instance, radial projection visualizations like Star Coordinates or RadViz [22, 23] can be utilized to show scatter plot clusters, trends or outliers by using the features as projection dimensions. Lehmann et al. [24, 25] introduced a visual guidance approach for those projection visualizations and proposed a generalization of both visualizations to achieve a higher degree of freedom for finding suitable projections. Furthermore, in [26] an experimental study compared the effectiveness of global features for ranking scatter plots by similarity.

2.4. Delineation of Our Approach and Novelty

Our work uses a feature-based approach for an interestingness ranking of scatter plots based on their contained local motifs. Other than previous approaches, which use global features, we here consider *local* properties of interest in scatter plots. Therefore, we complement global approaches. Our work is inspired by techniques from image processing and in particular the segmentation of local areas-of-interest in images and feature-based clustering. We employ the idea of a minimum spanning tree-based clustering, as introduced by Jana and Naik [27], to segment scatter plots into scatter plot patterns. To the best of our knowledge, this work is the first to apply the $tf \times idf$-scheme from information retrieval [28] for weighting and ranking scatter plot patterns.

3. Overview of Our Approach

The main goal of our approach is to guide the analyst through the exploration process, when facing a data set with a large number of individual scatter plots.

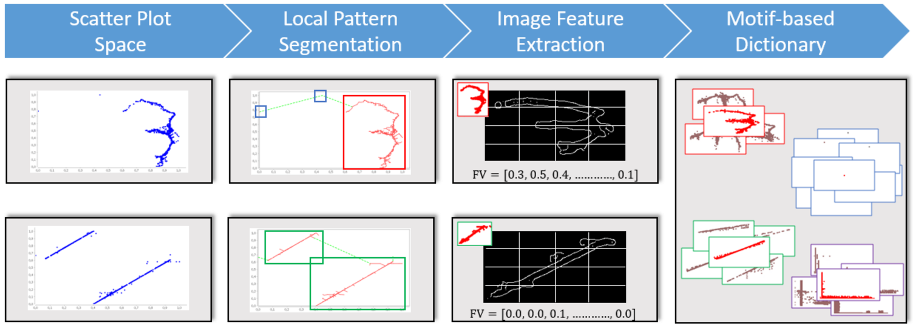

Scatter Plot Space
Local Pattern Segmentation
Image Feature Extraction
Motif-based Dictionary
FV = [0.3, 0.5, 0.4, …………, 0.1]
FV = [0.0, 0.0, 0.1, …………, 0.0]

Figure 1: Proposed analysis workflow to generate a motif-based dictionary and to derive a local interestingness score. The first step is to extract the local segments of each scatter plot by an adapted minimum spanning tree clustering approach (first and second columns). Then, we extract visual features of the individual local scatter plot segments (third column). These features are input to a subsequent clustering step done by applying k-means to the set of all segments. This leads to a number of k clusters, and for each cluster the medoid segment is chosen as the motif to represent the cluster (see last column).

Our main idea is to compute a *dictionary* of local scatter plot segments from the set of all scatter plots. The dictionary will contain prototype scatter plot segments (called motifs) that represent the different local scatter plot shapes occurring in a given data set. We form this dictionary by first partitioning all scatter plots into a set of local scatter plot segments. We then apply a clustering step that produces a number of clusters (i.e., dictionary entries), which represent the local motifs occurring in the overall data set. This clustering step relies in turn on visual features extracted from the individual local scatter plot segments. Based on a $tf \times idf$-analysis of the data using this dictionary, we compute interestingness scores for the individual scatter plots. Figure 1 shows our analysis workflow, which is detailed in the following.

Segmentation of Local Scatter Plot Patterns. The automatic segmentation of scatter plots is the basis of our interesting measure and hence requires special attention. Since each scatter plot may contain a distinct set of characteristics regarding its motifs (e.g., number, shape), its points (e.g., density) and the input scale of the dimensions, a flexible segmentation method is needed. A manual adjustment of segmentation parameters or the incorporation of domain knowledge in the segmentation process is practically infeasible, because many data sets under consideration contain possibly thousands or more plots. Therefore, the segmentation technique should be parameter-free and capable of finding motifs regardless of their shape. Since basic potential segmentation techniques like k-means or DBSCAN would not satisfy these requirements, we extended a minimum spanning tree (MST) based clustering technique. Another important prerequisite for the segmentation technique is to extract meaningful motifs that have a strong connection to the human perception. Experiments in [29] have shown that the MST method produces similar structures in the constellation of connection pairs of points (stars) as humans. The idea of the segmentation technique is to represent the data by means of a minimum spanning tree and iteratively remove the longest edge to derive an appropriate amount local scatter plot segments. Recent research on MST clustering has been conducted by Jana et al. in [27]. While their MST approach assesses the clustering quality in each iteration by an internal validity criterion [30], we follow-up on their research by introducing an *outlier-insensitive technique* that focuses on larger clusters containing more than one point. Basically, our approach can also be replaced by single linkage clustering analysis by processing the pairwise-distances in a certain order. The only drawback is that our outlier detection approach is not designed for this procedure and would fail at the initial step. Note that we discuss the impact of different quality criteria and their influence on the exploration process and the final result set in the following Sections. The final scatter plot segmentation is achieved by considering the clustering with the overall best assessment.

Dictionary Formation. After we have extracted local scatter plot segments, we compute the motif dictionary by clustering the set of local segments. To this end, we apply k-means clustering. This clustering requires an appropriate vector-based description of each segment. A recent study has shown that edge orientation and density features are effective to distinguish scatter plot shapes [26]. We therefore compute these features and feed them into a k-means clustering to produce the dictionary.

Interestingess Score and Visual Exploration. We compute a measure of interestingness for each scatter plot. To this end, we rely on the notion of $tf \times idf$-analysis from information retrieval. Briefly, we consider each entry in the dictionary (motif) as a visual word. Intuitively, a scatter plot which contains one or several instances of a motif (high term frequency), which does not occur in many other scatter plots (low document frequency), is considered important. We use this intuition to define a measure for ranking the interestingness of scatter plots. Also, given that we have segmentation and dictionary, we can apply color-coding to visualize the distribution of motifs across many scatter plots for interactive exploration (see also Section 6.1).

4. Global Interest-Measure based on Local Motifs

This section provides a technical overview of our implementation for detecting interesting scatter plot motifs and presents our aggregation scheme from local to global interestingness scores.

4.1. Automatic Motif Segmentation in Scatter Plots

We present an enhanced, parameter-free minimum spanning tree based clustering method. A core part of the method is the assessment of the clustering quality, typically defined by simultaneously achieving a high intra-cluster and low inter-cluster similarity. Following up on the research of Jana and Naik [27], we conduct a performance evaluation with different well-known internal validity scores (*F-Ratio* [31], *Inter-Intra Ratio* [32], *Davies-Bouldin Index* [33], *Silhouette Coefficient* [34]) to qualitatively discuss the choice of an appropriate clustering quality measure. Each measurement was applied together with the MST clustering technique to eight distinct pre-classified, ground truth scatter plots[1]. To test the segmentation approach, we chose scatter plots that differ

[1] Shape sets. Collected by the 'School of Computing, University of Eastern Finland' (`http://cs.joensuu.fi/sipu/datasets/`) Accessed 05/2015.

in the number of contained clusters, cluster shapes and cluster densities. The performance of the quality indices was measured by means of an external cluster evaluation measurement, the *Rand Index* [35]. In our experiments, the *F-Ratio* validity index outperformed the others with an average clustering accuracy of 89.3%. Due to this result, we employ the *F-Ratio* index as a clustering quality measure. Since the global clustering quality optimum would be inherently achieved after removing all edges from the MST and every point would constitute its own cluster we limit the number of overall clusters to $\lceil \sqrt{\frac{n}{2}} \rceil$ with n being the number of points [36].

By limiting the number of clusters to this threshold, another problematic issue arises: The removal of an edge, which connects an outlying point to the MST, can lead to a cluster of size one. By considering each cluster regardless of its content, the cluster limit might be reached prematurely. One naïve way to avoid this problem would be to simply ignore clusters of size one. Although this would eliminate the case, non-outlier points with only one connection within the MST would be discarded as well. Thus, the point should be taken into consideration by reducing the clustering quality score. In order to distinguish outliers from cluster points, an outlier detection method is applied if the removal of an edge results in a cluster of size one. We rely on a distance-based outlier detection method that considers the length of the last removed edge. If the edge length is high, compared to the edge lengths within an user-defined neighborhood area, then the outer vertex of the edge is marked as an outlier. Detected outliers are ignored in subsequent iterations of the algorithm.

For the construction of the minimum spanning tree, several algorithms are available (e.g., Kruskal's [37] or Prim's algorithm [38]). The MST algorithms take a connected, undirected graph for which we suggest to use the graph resulting from a Delaunay triangulation. This preprocessing step has the advantage to minimize the memory consumption to a linear level, which would otherwise be quadratic for complete graphs. Since the Delaunay triangulation is a supergraph of the MST, no relevant information is lost.

4.2. Dictionary-Based Interestingness Score

After we identified all sets of connected components (scatter plot segments), we group all similar segments in the scatter plot space and build a motif-based dictionary. The dictionary contains information about the distribution and frequency of segments, and is used to determine the local interesting score. Therefore, the characteristics of the segments need to be described by a suitable feature vector. While many different visual features are possible candidates, we decided to use gradient and density features, as these have been shown to work robust for global comparison of scatter plots [26].

To achieve this goal, we generate a normalized image for each local segment, scaled to the unit square. From these, we compute edge orientation and density features. Specifically, we subdivide the normalized image of a segment into a regular 16×16 grid and compute point density and a histogram of edge orientations for each cell. In order to robustly extract edge orientations, the segment images are blurred by applying a Gaussian filter and converted into an edge image with the help of Laplacian image filtering [39]. By means of these visual features, we cluster the segments regardless of position or axes scales. We apply a k-means clustering on the feature vectors of all local segments to form the motif dictionary, as illustrated in the last step of Figure 1. An essential step here is the parameter setting k for the number of dictionary entries, since it influences the quality of the dictionary and consequently the local interestingness score. To determine an appropriate setting for k, we developed a visual exploration front end for experimental tests (cf. Section 5.1), which visualizes the motif dictionary for different settings of k.

The set of clusters is the basis for computing the local interestingness score and expresses the uniqueness of a motif, and how discriminant the motif is regarding the entire scatter plot space. Accordingly, scatter plots containing overall rare and locally frequent motifs are ranked higher, and suggested to the analyst for inspection in an interactive system.

$$MU_{score}(q) = \frac{1}{|\{p \in Dict[q]\}|} \tag{1}$$

Equation 1 shows the proposed *Motif Uniqueness (MU)* score and how we measure the local interestingness for a given segment q. We divide one by the total number of segments p in the data set that belongs to the same motif q (i.e., the cluster size of the motif).

4.3. Global Interest Measure

The overall goal of our approach is to find interesting scatter plots for the exploration, containing discriminative local motifs. The global interest measure should reflect the interestingness of a given scatter plot based on the frequency of its local motifs in the entire scatter plot space. It is comparable to the text mining approach $tf \times idf$ [28], which uses the importance of a word to rank a document in a corpus. Instead of using the term frequency (*tf*), that computes the frequency of a term in a document, we use the *motif uniqueness* score from Section 4.2. It reflects how interesting and discriminant a motif is with respect to the corpus/scatter plot space. The basic idea of this local score is to weight frequent motifs (e.g., single dots or stripes) lower, and vice versa to weight discriminant motifs (e.g., complex patterns) higher.

The global interestingness measure is derived from these local factors in combination with an overall interestingness score. It corresponds to the inverse document frequency (*idf*) in text mining. The inverse document frequency is a measure to compute the overall importance of a term across all documents and follows the same idea as our second weighting factor that we call *inverse scatter plot frequency (ISPF)*. The difference to our approach is that we take the dictionary information and visual features into account and measure whether a motif is common or rare across all scatter plots. As shown in Equation 2, this score is obtained by dividing the total number of scatter plots N by the number of scatter plots sp containing one of the motifs in the dictionary cluster, and then taking the logarithm of that quotient. The substantial idea of this second weighting factor is to identify if a dictionary entry is based on many scatter plots containing such a motif, or e.g., just one scatter plot that contains many identical motifs.

$$ISPF_{score}(q) = \log \frac{N}{|\{sp \in Dict[q]\}|} \tag{2}$$

All local motif scores of a scatter plot are accumulated to produce the global interestingness score. Thus, scatter plots containing different and infrequent motifs achieve a higher score and are thereby considered as more interesting. Our proposed aggregation scheme for this interest measure is specified in Algorithm 1. For comparison reasons, we divide the aggregated global scatter plot interest score by the number of local motifs. Alternatively, analysts can use a range factor to prioritize the number of desired motifs and can penalize scatter plots containing more or less motifs. By means of this interest measure approach, we are able to automatically extract interesting views for the exploration of large scatter plot spaces.

Algorithm 1: Computation of a global interest measure

```
Input: $motifDict, S$
Result: List of global interest measures
foreach $scatterplot$ in $S(s_1, ...s_n)$ do
    $localMotifs$ = get motifs of $scatterplot$
    foreach $m$ in $localMotifs$ do
        $dictIndex$ = get dict index of $m$
        $localScore = MU(dictIndex) \cdot ISPF(dictIndex)$ $globalScore$ +=
          $localScore$
    $globalScore = globalScore/$size of $localMotifs$
    add $globalScore$ to $resultList$
return $resultList$
```

5. Visual Exploration

In this section, we introduce our visual exploration approaches and demonstrate how it support the selection of an appropriate dictionary size.

5.1. Identification of Similar Local Motifs

Selecting an appropriate dictionary size is difficult and has an impact on the subsequent process of finding local motifs and interesting global scatter plots. Especially for large and complex data, it is crucial to define a good cluster parameter k. Therefore, we developed a visual exploration tool to support analysts in the search process, find appropriate parameter settings, and finally suggest interesting scatter plots for exploration.

The tool involves a global overview in the form of a scatter plot matrix and a detailed dictionary view of all clustered motifs, as depicted in Figure 2. It allows the analysts to experiment with different clustering settings for a given data set. The dictionary view provides insights into the quality of the parameter setting and shows core information like cluster representatives and cluster size. The cluster size indicates the frequency of a particular representative motif in the scatter plot space. Moreover, it hints on the practicability of the chosen clustering parameter k. To represent the cluster, we chose the local segment, which is the nearest neighbor to the k-means prototype. By clicking on a dictionary entry, all cluster members of a motif will be highlighted in the linked scatter plot matrix. Conversely, it is possible to highlight all corresponding motifs by clicking on a given segment in the scatter plot matrix. Moreover, we distinguish by different color-codings the different motifs occurring in the data set. Thus, users can quickly recognize the distribution of individual motifs across a large scatter plot space. A further benefit of this overview is that users can estimate whether the cluster extraction threshold is configured appropriately, or whether the number of clusters should be increased or decreased.

5.2. Vector Space of Local Motifs

Besides the global scatter plot matrix, a projection visualization of all local scatter plot segments can be used to assess the dictionary quality. By means of this view, analysts may better understand the feature vector space of the local motifs and can readjust the settings to achieve a more reasonable dictionary. As a basis for the local motif projection view, a visualization technique is needed,

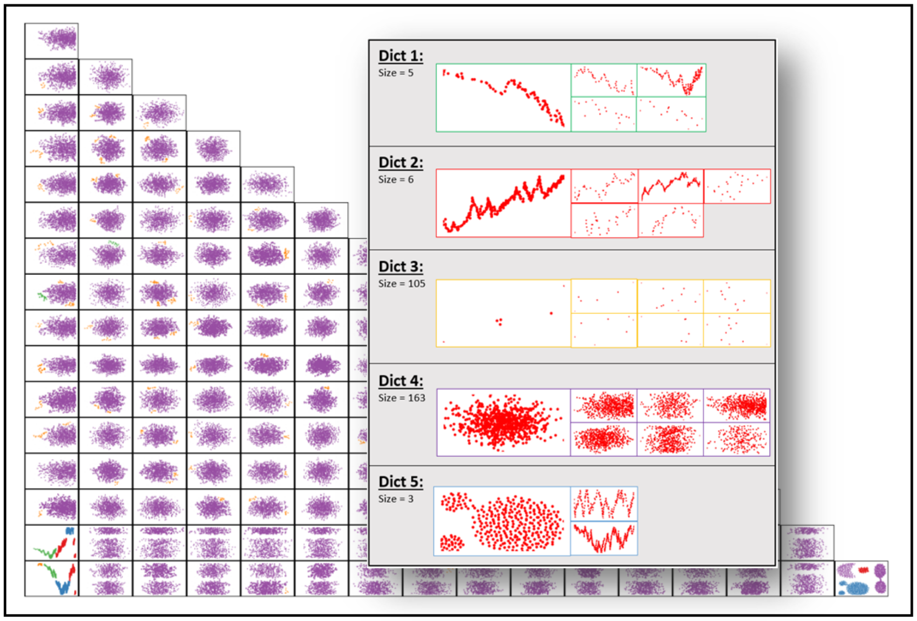


Figure 2: Scatter plot matrix overview of the synthetic data set and the resulting dictionary with five entries. The first 15 dimensions consist of Gaussian clusters and the last rows are combinations with the *Aggregation* data set. By means of the displayed dictionary view and the corresponding color coding, analysts can easily determine a good dictionary size for a given data set.

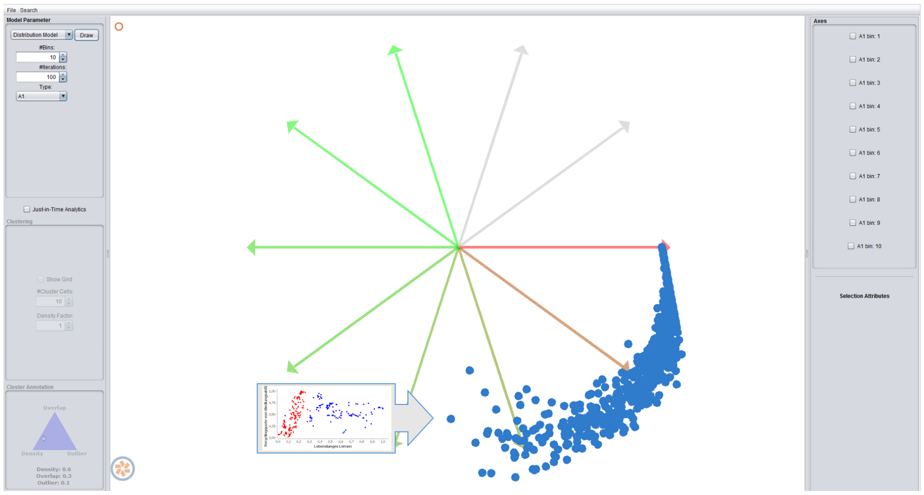
which is capable of displaying high-dimensional data. We chose the Star Coordinate visualization [22] as it fulfills this requirement and already provides sophisticated interaction mechanisms, such as axes rotation and modification, which support the understanding of high-dimensional spaces. Each point in the resulting visualization is a projection of the original feature vector and can be seen as a link back to a local pattern, as well as to the scatter plot it was originally taken from. The visualization itself uses a two dimensional plane on which axes are arranged in a circular manner - one axis for each dimension of the original data space. The axes have initially the same length and equal angles between each pair of them. In order to project a data point onto the two dimensional plane, the value of each dimension is mapped to the corresponding axis. Thereby, the highest value within a dimension is always mapped to the end of the axis while the smallest value is mapped to the origin in the middle of the plane. After that, these single mappings are linearly combined to achieve the data points final position. This can also be formalized mathematically by Equation 3,

$$p_j = \overrightarrow{O} + \sum_{i=1}^{n} \overrightarrow{a_i} \cdot \frac{d_{ji} - min_i}{max_i - min_i} \tag{3}$$

where p_j represents a data point $d_j = (d_{j1}, d_{j2}, ..., d_{ji}, ..., d_{jn}) \in D \subset \mathbb{R}^n$ after projecting it onto the 2D plane. Furthermore, $\overrightarrow{O}$ is the position vector of the origin, $\overrightarrow{a_i} \in A$ is a two dimensional axis vector with $A = \langle \overrightarrow{a_1}, \overrightarrow{a_2}, ..., \overrightarrow{a_i}, ..., \overrightarrow{a_n} \rangle$ and min_i and max_i are the minimal and maximal values of dimension i.

The user is able to modify this projection and hence the resulting view by modifying one or more of the axes. An axis can be dragged resulting in a change of its direction and length. A modification immediately updates the visualization and enables the user to follow the re-projected points. While points which are projected to zero in a dimension are not influenced by a modification of the respective axis, higher values will result in a faster movement. Thereby, the user establishes a natural grasp of the data values and their dependency regarding the respective dimension. Another benefit of this visual effect is that overdrawing points become visible by modifying the dimension axes. A demonstration of the projection visualization is shown in Figure 3.

Since we use a visual way of exploring the feature vectors, it is also essential to have access to the original local patterns to which points in the visualization refer. To that end, an user can hover the mouse cursor over any point and the respective scatter plot from which the pattern was originally taken shows up in form of a tooltip. The tooltip depicts the local pattern colored in red while the remaining points of the scatter plot are displayed in blue. Furthermore, points can be selected in order to get additional information about the underlying data values or their distribution. Selecting a point will also highlight sibling patterns which means that all local patterns originating from the same scatter plot will be marked. This enables the user to compare the local contents of a single scatter



File Search

Model Parameter

Distribution Model ▾ Draw

#Bins: 10

#Iterations: 100

Type: A1

☐ Just-in-Time Analytics

Clustering

☐ Show Grid

#Cluster Cells: 10

Density Factor: 1

Cluster Annotation

Density: 0.6
Overlap: 0.3
Outlier: 0.1

Axes

☐ A1 bin: 1
☐ A1 bin: 2
☐ A1 bin: 3
☐ A1 bin: 4
☐ A1 bin: 5
☐ A1 bin: 6
☐ A1 bin: 7
☐ A1 bin: 8
☐ A1 bin: 9
☐ A1 bin: 10

Selection Attributes

Figure 3: Initial Star Coordinate projection by visualizing the local motifs of the *eurostat* data set (see more about the data in Section 6.2). Each point represents one local scatter plot segment, which is highlighted in red within a tooltip. All points in the projection view are mainly influenced by the four dominating axes (colored in red, orange and dark green) and thus are located at the bottom right corner.

plot and helps to evaluate whether the scatter plot consists of very different or rather similar patterns. Figure 4 shows the selection of a local pattern and its corresponding auxiliary visualization in form of the original scatter plot.

Relevance Score: With an increasing number of feature vector dimensions, also the number of corresponding axes grows accordingly. Hence, in order to create an informative view, the user has to select and modify one or more axes from a big number of potential candidates. Choosing the right axes is so far not practicable since the user cannot predict the visual impact of the different axes until he modified each of them manually. In this context, Chen recently studied several Star Coordinate visualization models and described a heuristic rule how to effectively work with this type of visualization [40]. Chen differentiates between ”visually dominating dimensions” dimensions (axes) whose modification result in a significant change of the view and ”fine-tuning dimensions” dimensions that have no or only a small influence on the visualization.

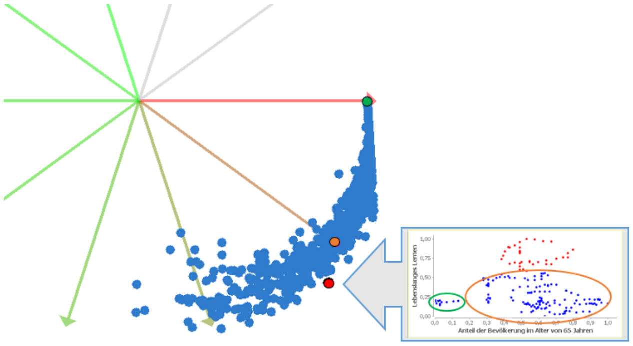


Figure 4: Illustration of a local pattern selection. The red point depicts the selected scatter plot segment and is highlighted in the corresponding tooltip (red points). By selecting a local pattern, all sibling patterns will also be highlighted in the projection view (orange and green points).

Chen's suggestion is to start modifying axes corresponding to "visually dominating dimensions" and use axes of "fine-tuning dimensions" to finalize the visualization. Since this heuristic seems reasonable, we follow Chen's suggestion and introduce a relevance score to identify these types of dimensions. We consider the distances between differing points within each dimension and calculate an average distance to determine the score. A high average inter-point distance will thereby lead to a notable change of the visualization while dimensions solely consisting of very similar values will show a homogeneous movement of all points (i.e., every point will move almost the same amount in the same direction). Let $S_d = (s_{d1}, s_{d2}, ..., s_{dn})$ be an ascending ordered sequence of real numbers corresponding to the feature vector values of one particular dimension d, then R_d denotes d's relevance score defined by Equation 4.

$$R_d = \frac{1}{n-1} \sum_{i=1}^{n-1} s_{d(i+1)} - s_{di} \tag{4}$$

Figure 3 shows the relevance score R_d mapped to the color of the corresponding axis. Red indicates a high relevance value and hence a visually dominating dimension, green indicates a "fine-tuning dimension" and gray displays a dimension that has no visual influence at all.

Figure 5 exemplifies a modification of the initial Star Coordinate visualization (seen in Figure 3), in which we spread out the four most dominating axes to the corners and decrease the length of all other axes that have little or no influence on the projection space. Thus, the projection space can be fully exploited by using the most meaningful axes and simplifies the detection of similar local pattern groupings. By taking a closer look at the highlighted scatter plot patterns, we can already clearly identify separated groupings of local patterns in the projection space. At the lower region of the projection space are patterns located, which contain wide-spread motifs (blue tooltips), skinny linear motifs are located left (purple tooltips), and motifs including small dense point clouds are located at the top (green tooltips). By just considering the motifs at the outer rim of the projection space, we are able to identify three different groupings.

6. Application of Motif-based Dictionary

We now demonstrate the usefulness of our interest measure and the global scatter plot ranking by means of our visual exploration tools. First, we use a synthetic data set as a proof-of-concept to showcase our proposed interest measure. We then make use of the interest measure on a real-world data set and explore the suggested scatter plots.

6.1. Synthetic Data: Interestingness Measure

We created a synthetic data set by merging 15-dimensional Gaussian clusters with the two-dimensional *Aggregation* data set presented in [41]. Since, the aggregation data set consists of a small sample size (788 records), we randomly created Gaussian clusters with the same size and merged the data, as illustrated in the background of Figure 2. The original scatter plot of the aggregation data set is located at the bottom right corner of the matrix. The experiment was designed to depict that motifs of the Gaussian dimensions (purple motif), which

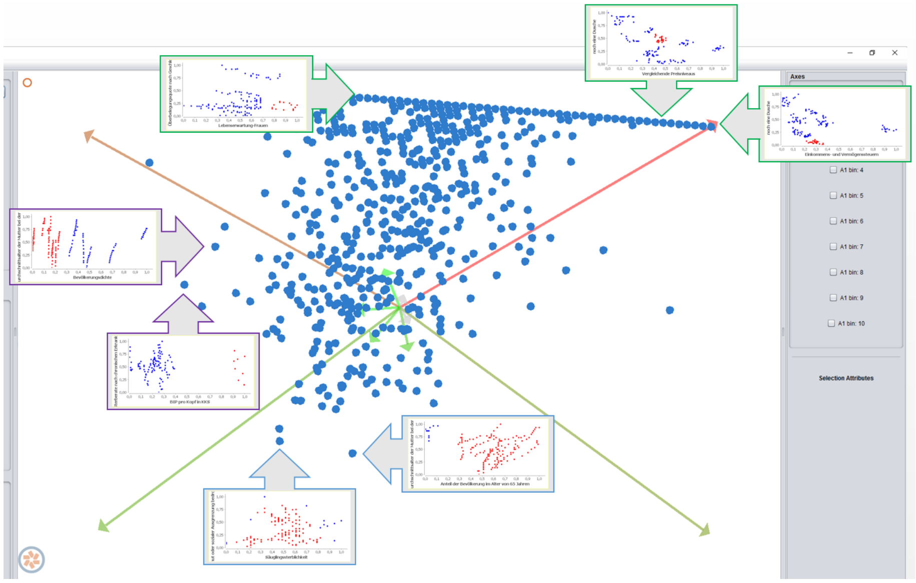
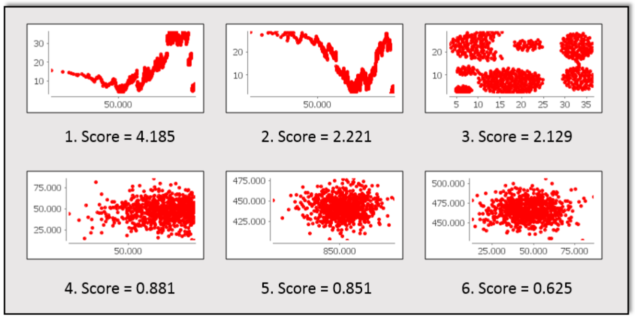


Figure 5: Modification of our initial Star Coordinate projection view (shown in Figure 3). The four dominating axes are widely spread to the corners in order to reduce overdrawing points and simplify the detection of similar local pattern groupings. Different motif types are spatially divided in the projection space.

appear more often will also result in a low local and overall interestingness score. In contrast, scatter plots that were merged with one of the aggregation data dimensions (last two rows) contain more complex and outstanding motifs, and will thus be rated more interesting.

The first step of our approach is to determine the interesting scatter plot segments by running our adapted MST approach (cf. Section 4.1). After the segmentation step, we have extracted 282 local segments from 136 scatter plots. When looking at the scatter plot matrix, we can see that the data set contains only a few kinds of different motifs. In this case, we recommend choosing a small k (e.g., between three and five) to keep the quality of the dictionary high and clearly separate the different motif shapes. Choosing a too large dictionary size would lead to splitting up the homogeneous motifs of the Gaussian clusters into several dictionary entries and thus will distort the local interestingness score. On the contrary, a too small dictionary size will merge dissimilar motifs and

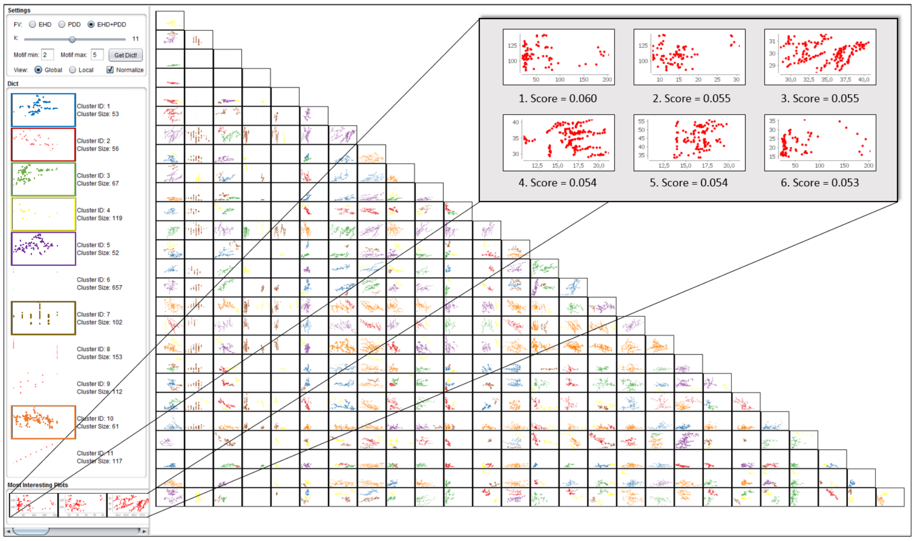


1. Score = 4.185
2. Score = 2.221
3. Score = 2.129
4. Score = 0.881
5. Score = 0.851
6. Score = 0.625

Figure 6: The six most interesting scatter plots of the synthetic data set for exploring local motifs. The global scores are obtained by aggregating all local motif scores ($MU \times ISPF$) of a given scatter plot.

also negatively influence the ranking.

In the experiment, depicted in Figure 2, we found a good dictionary setting by using the combined image descriptor, which takes the edge orientation and density of a motif into account (cf. Section 4.2) and chose a dictionary size of five. Thus, we received a dictionary with five well-separated clusters, containing a negative trend motif (green), positive trend motif (red), sparse point clouds (orange), dense point clouds (purple) and a motif cluster with wide-spread distributions (blue). The largest motif cluster is represented by the purple color with 163 similar segments, followed by the orange cluster with 105 segments. As one can see, all patterns are highly similar in the scatter plot space except those from the original aggregation data set and the two scatter plots in combination with the first dimension. Consequently, our interest measure ranks scatter plots with the purple and orange motifs less interesting than the other motif groups. A result overview of the six top-ranked scatter plots is shown in Figure 6. As expected, the three scatter plots containing deviant motifs achieve the highest global scores. The reason why the first three scatter plots received significantly higher scores is due to their higher ranked local motifs. The original aggregation scatter plot has been ranked third after the two line shaped scatter plots, because of the poor local score of the two purple motifs. Very unexpected was the incident that all other scatter plots derived from the two aggregation data


Settings

FV: ○ EHD ○ PDD ● EHD+PDD

k: 11

Motif min: 2 Motif max: 5

View: ● Global ○ Local ☑ Normalize

Dict

Cluster ID: 1
Cluster Size: 53

Cluster ID: 2
Cluster Size: 56

Cluster ID: 3
Cluster Size: 67

Cluster ID: 4
Cluster Size: 119

Cluster ID: 5
Cluster Size: 52

Cluster ID: 6
Cluster Size: 657

Cluster ID: 7
Cluster Size: 102

Cluster ID: 8
Cluster Size: 153

Cluster ID: 9
Cluster Size: 112

Cluster ID: 10
Cluster Size: 61

Cluster ID: 11
Cluster Size: 117

Figure 7: Our visual exploration tool for global and local scatter plot analysis. By means of this tool, analysts can derive different motif-based dictionaries by adjusting the parameter settings and thus achieve various interesting scatter plots suggestions for exploration. The parameter settings, dictionary view and the resulting global interest ranking are located on the left hand side. Local motifs of a given setting can be highlighted in the scatter plot matrix to assess the dictionary quality.

dimensions are not in the top six results. This can be explained by the fact that scatter plots on rank four to six also contain the higher ranked motifs of the orange and green dictionary entry.

6.2. Real-World Data: Interestingness Analysis

The second evaluation data set is retrieved from the *eurostat*[2] data repository. The data repository provides in total 5500 data sets each containing information about a European related topic, such as economy, population and industry. We extracted a data subset containing 27 statistical attributes from 28 EU countries that show temporal changes over the last decades. We created

[2] Statistical Office of the European Union (`http://ec.europa.eu/eurostat`). Accessed 02/2016.

a scatter plot matrix with 351 unique scatter plots from these 27 dimensions in which each data instance (point) represents one country at a specific year. The corresponding scatter plot matrix is illustrated in Figure 7.

As in the previous example, we start the interestingness search process with segmenting the scatter plot space into local segments and select a good setting for the dictionary. Our segmentation approach returned 1549 local segments of the 351 scatter plots. We are using the combined image descriptor for characterizing the motifs in this experiment. By exploring the projection space of the local motifs (see Figure 5), we first roughly estimate the dictionary size between 8 and 14. Then we iteratively highlight the most considerable motifs in the scatter plot matrix to identify the similarity of a dictionary entry and thus proof the quality of the settings. Finally, we decided to choose a dictionary size of 11 for further analysis. As Figure 7 depicts, one can clearly recognize the dependencies between similar motifs and the dimensions in the scatter plot matrix. For instance, if we consider the brown motif class (dictionary cluster ID 7), we are able to identify all the dictionary items in column two, four and five. The same applies to the orange motif class with sparse negative trend direction (dictionary cluster ID 10), which are mostly located in row 16 and 18. Finding such properties in the scatter plot matrix may lead to first insights in the local motif analysis.

The top ranked scatter plots of our chosen setting are outlined at the bottom left corner of our visual exploration tool. An enlarged excerpt of the best six rankings is also shown in Figure 7. On closer inspection, we can see that all suggested scatter plots contain significant motifs, which may be interesting to analyze. As an analysis example, we want to focus on the scatter plot ranked on the third place. The scatter plot shows separated motifs with several positive trend directions shifted on both axes. These motifs describe the relation between the mean duration of working life against the mean age of women at childbirth of the population in all EU countries. It becomes clear that the total work duration of women decreases when they become a mother earlier and for some reason, can no longer work. Additionally, it would be interesting to analyze these different groupings in relation to other non-numerical attributes, such as geolocation and see which countries share similar characteristics and how they change over time.

7. Discussion of Limitations and Extensions

In our approach, we are interested to guide the exploration of scatter plots based on the notion of interestingness of local motifs in large scatter plot spaces. The concept of local pattern analysis is novel in that it extends beyond most feature-based scatter plot analysis methods, which consider global scatter plot features. Our solution is a first step to extend the analysis for local scatter plot patterns and depends on the choice of methods applied.

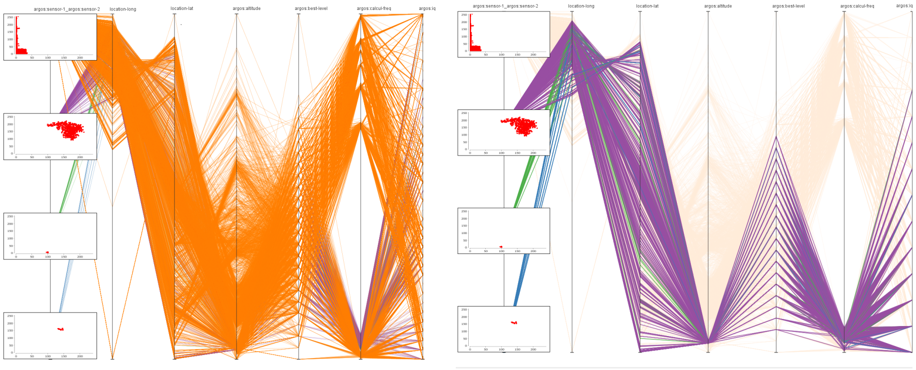
Regarding the segmentation, many alternatives could be considered. First, data space segmentation approaches are possible, but also, image-based segmentation approaches could be considered. While our implemented approach is based on partitioning distributions in the data space, other data space segmentation approaches are possible. For example, a regression tree can be learned for finding a (non-)linear partitioning of the scatter plot space. Alternatively, a wide range of options from the image analysis community are available and not yet explored on scatter plots. We are planing to experiment with convex hull calculations on the rendered scatter plot images to find local motifs. One advantage of this approach is that data space axis ranges are normalized by definition, such that their treatment is not important anymore. In our case, we normalize each scatter plot segment for the unit square. While this normalization supports easy extraction and comparison of features, it ignores different scales, aspect ratios and spatial relations among the patterns. All of this could be taken into account by extending of the feature vector, based on the application need.

Then, we are planning to integrate further motif descriptors in our system. While currently, an edge-descriptor and a density-based descriptor proved useful for our case studies, we expect that a greater variety of scatter plot motifs can be described with other feature descriptors. One option is to apply Hough analysis to detect presence of basic shapes in scatter plots such as lines or areas and form a descriptor from these. Other than that, regressional features, such as described in [20], could be integrated. One advantage of the latter descriptor is that it can be interpreted in terms of a regression model, where as the density and edge orientation features used here, are low-level and cannot be easily interpreted by the user. These descriptors will bring an advanced semantics level in the motif detection, which needs to be understood and researched more extensively.

Indeed, a key design issue is the definition of an interestingness function based on motif distributions. Our proposed score is basically representing the notion of outlyingness (or sparsity). However, many other notions of interestingness could be defined, based on the motif distribution. For instance, a distribution where specific combinations of motifs in a scatter plot occurring frequently, could be valuable.

We exploit the motif analysis and ranking in an interactive scatter plot matrix representation, which allows users to compare and overview the different motifs by color-coding. The visual exploration tool could be enhanced in several ways. An overlay of cluster members with semi-transparent drawing may be effective to this end. So far, we draw each local pattern in its original shape, showing the cluster (motif) membership by color coding. An alternative that could scale better for large scatter plot spaces, might be to abstract the shape of each motif and replace each occurrence of a motif in a scatter plot by its simplified version. The recently proposed so-called Visual Guidance Pictograms [24] could be a starting point to this end.

Finally, one might improve the visual exploration tool by including additional non-numerical attributes, such as geolocation or textual data, to gain further information about the extracted motifs. Thus, for instance, it could be examined why certain motifs differ in their shapes or similar motifs are shifted along a scatter plot axis. Another idea would be to extend other existing visualization techniques for motif analysis. The traditional parallel coordinate plot could be extended with visual motif axes to analyze the extracted motifs along other numerical or categorical axes. We already have taken the first step into



(a) Without brushing interaction.

(b) Brushing on the three lower motif.

Figure 8: Parallel coordinates visualization with extended visual axes. All data records are displayed with different color mappings according to the local motifs.

this direction and implemented a prototype visualization, which connects visual motifs with other numerical attributes inside a parallel coordinates visualization, as shown in Figure 8. We have taken an scatter plot containing a high interest score and extracted its four local pattern on a visual axis of the parallel coordinates plot. The two lower motifs represent small point clouds that are only related to a few data records (polylines), whereas the two upper motifs represent large significant patterns and are also connected to many data records in the parallel coordinates plot. By just considering the polylines of the two upper motifs (Figure 8 (a) - orange and Figure 8 (b) - purple), we are already in the position to discover dimensional dependencies based on the extracted motifs.

8. Conclusion

We introduced a novel workflow in which we analyze the interestingness of automatically extracted local motifs to guide the exploration in scatter plot data. To assess the overall interestingness, we adapted the $tf \times idf$ scheme from information retrieval to the domain of scatter plot motifs. We derive the interestingness of local scatter plot motifs based on its occurrence among and within the scatter plot space. Furthermore, we developed interactive visual exploration tools with brushing and linking that supports analysts to find appropriate motif dictionaries and suggests interesting scatter plots for exploration. Finally, we applied the workflow on a synthetic and real-world data set to demonstrate how it can efficiently lead to interesting discoveries of local motifs. Our approach is only a first step into the direction of local analysis in large scatter plot spaces, and we have discussed a range of extensions to be done in future work.

Acknowledgment

This work was supported by the State of Baden-Württemberg within the research project Visual Search and Analysis Methods for Time-Oriented Annotated Data, With Applications to Research and Open Data.

References

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

[2] L. Wilkinson, A. Anand, R. Grossman, Graph-theoretic scagnostics, in: In Proceedings of the IEEE Symposium on Information Visualization, 2005, pp. 157–164.

[3] M. Sips, B. Neubert, J. P. Lewis, P. Hanrahan, Selecting good views of high-dimensional data using class consistency, in: Computer Graphics Forum, Vol. 28, Wiley Online Library, 2009, pp. 831–838.

[4] A. Tatu, G. Albuquerque, M. Eisemann, P. Bak, H. Theisel, M. Magnor, D. Keim, Automated analytical methods to support visual exploration of high-dimensional data, Visualization and Computer Graphics, IEEE Transactions on 17 (5) (2011) 584–597.

[5] J. Yang, Y.-G. Jiang, A. G. Hauptmann, C.-W. Ngo, Evaluating bag-of-visual-words representations in scene classification, in: Proceedings of

the International Workshop on Workshop on Multimedia Information Retrieval, MIR '07, ACM, New York, NY, USA, 2007, pp. 197–206.

[6] W. Cleveland, The shape parameter of a two-variable graph, Journal of the American Statistical Association 83 (402) (1988) 289–300.

[7] J. Talbot, J. Gerth, P. Hanrahan, Arc length-based aspect ratio selection., IEEE transactions on visualization and computer graphics 17 (12) (2011) 2276–82.

[8] M. Fink, J.-H. Haunert, J. Spoerhase, A. Wolff, Selecting the aspect ratio of a scatter plot based on its delaunay triangulation, IEEE transactions on visualization and computer graphics 19 (12) (2013) 2326–35.

[9] R. Motta, R. Minghim, A. de Andrade Lopes, M. C. F. Oliveira, Graph-based measures to assist user assessment of multidimensional projections, Neurocomputing 150, Part B (0) (2015) 583 – 598.

[10] A. Mayorga, M. Gleicher, Splatterplots: Overcoming overdraw in scatter plots, IEEE Transactions on Visualization and Computer Graphics 19 (9) (2013) 1526–1538.

[11] H. Chen, W. Chen, H. Mei, Z. Liu, K. Zhou, W. Chen, W. Gu, K.-L. Ma, Visual abstraction and exploration of multi-class scatterplots, IEEE Transactions on Visualization & Computer Graphics 20 (12).

[12] M. Sedlmair, A. Tatu, T. Munzner, M. Tory, A taxonomy of visual cluster separation factors, Computer Graphics Forum (Proc. EuroVis 2012) 31(3) (2012) 1335–1344.

[13] M. Sips, B. Neubert, J. P. Lewis, P. Hanrahan, Selecting good views of high-dimensional data using class consistency, Computer Graphics Forum (Proc. EuroVis 2009) 28 (3).

[14] D. J. Lehmann, G. Albuquerque, M. Eisemann, M. Magnor, H. Theisel, Selecting coherent and relevant plots in large scatterplot matrices, Computer Graphics Forum 31 (6) (2012) 1895–1908.

[15] S. Bremm, T. von Landesberger, J. Bernard, T. Schreck, Assisted descriptor selection based on visual comparative data analysis, Wiley-Blackwell Computer Graphics Forum 30 (3) (2011) 891–900, proceedings of Eurographics/IEEE-VGTC Symposium on Visualization 2011).

[16] G. Albuquerque, M. Eisemann, D. J. Lehmann, H. Theisel, M. A. Magnor, Quality-based visualization matrices., in: VMV, 2009, pp. 341–350.

[17] A. Anand, L. Wilkinson, D. T. Nhon, Visual pattern discovery using random projections, in: IEEE VAST, 2012, pp. 43–52.

[18] A. Tatu, F. Maaß, I. Färber, E. Bertini, T. Schreck, T. Seidl, D. A. Keim, Subspace Search and Visualization to Make Sense of Alternative Clusterings in High-Dimensional Data, in: Procedings of IEEE Symposium on Visual Analytics Science and Technology (VAST), IEEE CS Press, 2012, pp. 63–72.

[19] N. Elmqvist, P. Dragicevic, J.-D. Fekete, Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation, IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis 2008) 14 (6) (2008) 1141–1148.

[20] M. Scherer, J. Bernard, T. Schreck, Retrieval and exploratory search in multivariate research data repositories using regressional features, in: Proc. ACM/IEEE Joint Conference on Digital Libraries, 2011, pp. 363–372.

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

[22] E. Kandogan, Visualizing multi-dimensional clusters, trends, and outliers using star coordinates, in: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '01, ACM, New York, NY, USA, 2001, pp. 107–116.

[23] P. Hoffman, G. Grinstein, K. Marx, I. Grosse, E. Stanley, Dna visual and analytic data mining, in: Visualization '97., Proceedings, 1997, pp. 437–441.

[24] D. Lehmann, F. Kemmler, T. Zhyhalava, M. Kirschke, H. Theisel, Visualnostics: Visual guidance pictograms for analyzing projections of high-dimensional data, Computer Graphics Forum (Proc.EuroVis) 34 (3).

[25] D. Lehmann, H. Theisel, General projective maps for multidimensional data projection, Computer Graphics Forum (Proc.Eurographics).

[26] M. Scherer, T. von Landesberger, T. Schreck, A Benchmark for Content-Based Retrieval in Bivariate Data Collections, in: Proc. Int. Conference on Theory and Practice of Digital Libraries, 2012.

[27] P. Jana, A. Naik, An efficient minimum spanning tree based clustering algorithm, in: Proc. Int. Conference on Methods and Models in Computer Science, 2009.

[28] G. Salton, M. J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, Inc., New York, NY, USA, 1986.

[29] M. Dry, D. Navarro, A. Preiss, M. Lee, The perceptual organization of point constellations, 2009.

[30] Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu, Understanding of internal clustering validation measures, in: Proceedings of the 2010 IEEE International Conference on Data Mining, ICDM '10, IEEE Computer Society, Washington, DC, USA, 2010, pp. 911–916.

[31] L. A. Garcia-Escudero, A. Gordaliza, A. Mayo-Iscar, C. Matran, A robust maximal f-ratio statistic to detect clusters structure, in: Communications in Statistics - Theory and Methods, 2009, pp. 682–694.

[32] S. Ray, R. Turi, Determination of number of clusters in k-means clustering and application in colour image segmentation, in: Proceedings of the 4th

International Conference on Advances in Pattern Recognition and Digital Techniques (ICAPRDT'99), Narosa Publishing House, New Delhi, India, 1999, pp. 137–143.

[33] D. L. Davies, D. W. Bouldin, A cluster separation measure, Pattern Analysis and Machine Intelligence, IEEE Transactions on PAMI-1 (2) (1979) 224–227.

[34] P. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1) (1987) 53–65.

[35] W. Rand, Objective criteria for the evaluation of clustering methods, Journal of the American Statistical Association 66 (336) (1971) 846–850.

[36] K. V. Mardia, J. T. Kent, J. M. Bibby, Multivariate Analysis (Probability and Mathematical Statistics), Academic Press; 1 edition (January 27, 1976), 1976.

[37] J. Kruskal, On the shortest spanning subtree and the traveling salesman problem, in: Proceedings of the American Mathematical Society., 1956, pp. 45–50.

[38] R. C. Prim, Shortest connection networks and some generalisations, Bell System Technical Journal 36 (1957) 1389–1401.

[39] D. K. Park, Y. S. Jeon, C. S. Won, Efficient use of local edge histogram descriptor, in: Proceedings of the 2000 ACM workshops on Multimedia, ACM, New York, NY, USA, 2000, pp. 51–54.

[40] K. Chen, Optimizing star-coordinate visualization models for effective interactive cluster exploration on big data, Intell. Data Anal. 18 (2) (2014) 117–136.

[41] A. Gionis, H. Mannila, P. Tsaparas, Clustering aggregation, ACM Transactions on Knowledge Discovery from Data 1 (1) (2007) 4–es.