ScagExplorer: Using Scagnostics to cluster huge datasets

Tuan Nhon Dang*
University of Illinois at Chicago

Leland Wilkinson
University of Illinois at Chicago.

Abstract

We introduce a method for guiding interactive exploration of high-dimensional data. The method is based on nine characterizations of the 2D distributions of orthogonal pairwise projections on a set of points in multidimensional Euclidean space. These characterizations include measures such as, density, skewness, shape, outliers, and texture. Using with these measures, we can quickly generate a comprehensive summary of the 2D relations of variables in a large dataset with more than a hundred dimensions.

1 Introduction

Previous researchers have applied data-driven approaches to moderate-sized collections of scatterplots [8, 3]. The scale of these efforts has been constrained by display limits and computational complexity. In this poster, we have developed an alternative approach in order to deal with the scalability problem for SPLOMs in terms of data sets larger than one hundred dimensions. Our goal is to be able to organize these plots into meaningful subsets in reasonable time and to present these plots to users in a rich exploratory environment.

Our contributions in this poster are: We have proposed a way to retrieve similar scatterplots to a plot of interest-based euclidean distance in scagnostics space. And we have developed a dynamic algorithm that leverages force-directed graph methods to cluster the leader scatterplots which are obtained from the leader algorithm [4].

2 Related Work

2.1 Scagnostics

Scagnostics were introduced by Tukey and Wilkinson. The scagnostics measures depend on proximity graphs that are all subsets of the Delaunay triangulation: the minimum spanning tree, the alpha complex, and the convex hull. The graph-theoretic implementation of scagnostics are described in details in [8].

2.2 Feature-based Approaches

Seo and Shneiderman [7] computed statistical summaries (means, standard deviations, etc.) on univariate and bivariate distributions and then ranked them in order to identify similar distributions. Other researchers have developed scagnostics-type measures for parallel coordinates, pixel displays, and other graphics [2, 6].

Yang et al.[9] proposed a *Value and Relation* (VaR) technique explicitly conveying the relationships among the dimensions of a high dimensional dataset based on the data values in each dimension. The VaR technique helps users grasp the associations among dimensions. However, this technique fails to capture more complicated relations that can be captured by ScagExplorer.

TimeSeer [1] uses scagnostics for organizing multivariate time series and for guiding interactive exploration through high-dimensional data. TimeSeer consists of 2 systems: a SPLOM viewer and a time series viewer. The SPLOM viewer provides a guidance for selecting interesting pairs of variables. Then, the time series viewer can graph scagnostics time series of selected pairs of variables.

*e-mail: tdang@cs.uic.edu

3 ScagExplorer Components

This section explains our approach in detail. Figure 1 shows a schematic overview:

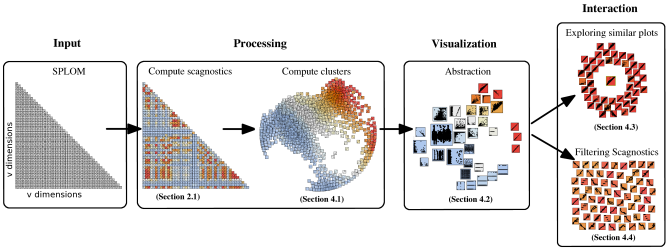

Figure 1: Schematic overview of ScagExplorer.

1. **Processing:** Our approach computes nine scagnostics measures of each scatterplot in the input SPLOM. Then, scatterplots are clustered based on scagnostics space.

2. **Visualization:** The leader plots in each cluster are displayed in the force-directed layout.

3. **Interaction:** Users can select a leader to see all similar plots in that cluster or filter scatterplots by their features.

3.1 Dissimilarity of Two Scatterplots

The dissimilarity of two scatterplots (S and P) is computed by the following equation:

$$Dissimilarity(S,P) = \sqrt{\sum_{i=1}^{9} W_i(S_i - P_i)^2} \qquad (1)$$

where S and P are two arrays of nine scagnostics of the two scatterplots and W_i is the weight of each scagnostics measure (user input).

3.2 Clustering Algorithm

The scatterplot matrix is a useful tool for displaying the correlation between a pair of variables. However, it is easy to run out of space as the number of variables increases. There are several solutions to deal with this scalability problem including dimension reduction, lensing [1], and SPLOM reordering [8, 5]. We use the leader algorithm [4] to cluster scatterplots. The complexity of this algorithm is $O(p)$ or $O(v^2)$ compared to $O(v^3)$ of Lehmann's reordering algorithm (where p is the number of scatterplots and v is the number of dimensions).

3.3 Displaying the leader scatterplots

After having clusters and their leader plots, we now use the force-directed layout to place them on a 2D view. The advantages of force-directed layouts are intuitiveness, flexibility, and interactivity. The main disadvantage is high running time. Since for every plot,

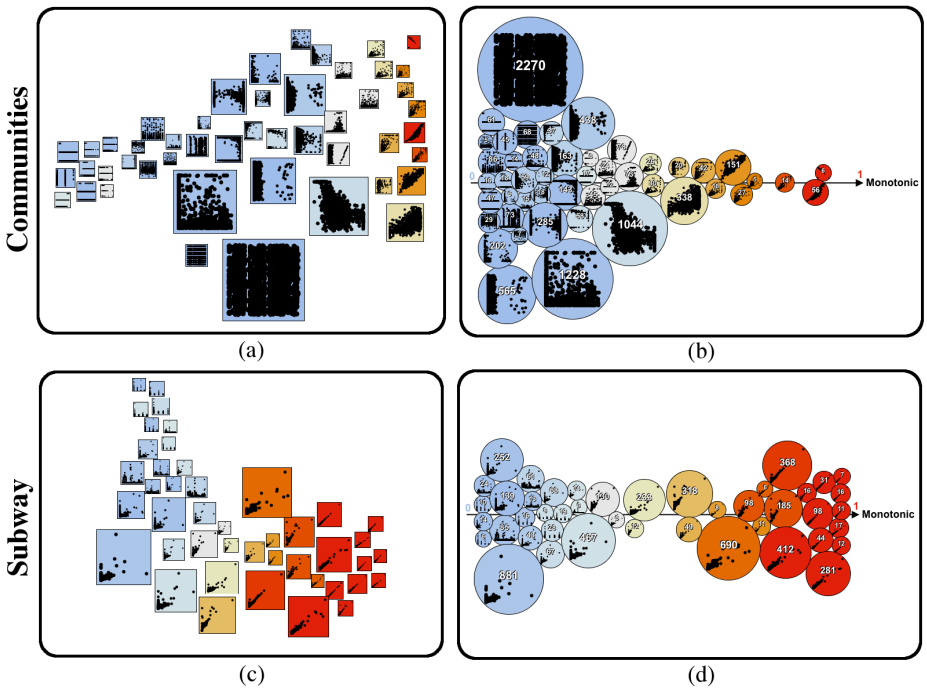

Figure 2: Visualizing the Communities and Subway. Left frames: overview layout. Right frames: Leader plots are aligned on X-axis (Monotonic).

we have to compute the attraction or repellant against all other plots, the running time at each iteration is $O(l^2)$ (where l is the number of leader scatterplots). However, l is much smaller than p and we can control the clustering algorithm so that it generates a reasonable numbers of leader plots (less than 100 leader plots). This running time is not remarkable.

Figure 2 shows how we display the leader plots of two different datasets in the forced-directed layout. The Communities data set has 128 dimensions (8128 scatterplots, each scatterplot contains 1994 data points). The Subway data set has 104 dimensions (5356 scatterplots, each scatterplot containing 423 data points). In particular, the frames on left summarize thousands of scatterplots in each dataset. The size of each leader plot is computed based on its cluster size (or the number of scatterplots in each cluster). The Kelvin temperature color scale is adopted to highlight Monotonicity (red plots are high Monotonic). In the frames on right, we aligned the leader plots on X-axis based on their Monotonicity. This alignment reveals the density distribution of scatterplots on Monotonicity. Notice that we have requested to display the number of scatterplots in each cluster on top of each leader plot. We also use circles instead of rectangles to present scattetplots. This option creates a better effect on displaying density compared to rectangle shapes. The forced-directed layout makes transitions from left frames to right frames smoothly.

4 Conclusions

ScagExplorer is a visual tool for analyzing high-dimensional datasets. The benefit of our approach is the compression we achieve by collapsing similarity searches from $O(n)$ to $O(1)$ through the use of scagnostics. The associated plots of the clusters are coherent. That is, they are directly comparable with each other in terms of similar shapes/distributions since we measure the relevant plots by inspecting all of their scagnostics. Instead, other approaches [5, 7] look at one measure at a time. Moreover, we implemented a quick clustering algorithm [4] with the complexity $O(v^2)$ where v is the number of variables in the data. This provides ScagExplorer with the scalability to handle huge datasets with hundreds of dimensions.

But there is also a drawback of the approach: one problem of the forced-directed layout is that random initial layout would yield different clusters of coherent plots. Therefore, different runs end up with different configurations. However, the final configurations of the same data are consistent because relevant leaders are grouped together. This provides a comprehensive summary of the input data.

References

[1] T. N. Dang, A. Anand, and L. Wilkinson. Timeseer: Scagnostics for high-dimensional time series. *IEEE Transactions on Visualization and Computer Graphics*, 99(PrePrints), 2012.

[2] A. Dasgupta and R. Kosara. Pargnostics: Screen-space metrics for parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics*, 16:1017–2626, 2010.

[3] S. Guha, P. Kidwell, R. Hafen, and W. S. Cleveland. Visualization databases for the analysis of large complex datasets. *Journal of Machine Learning Research - Proceedings Track*, 5:193–200, 2009.

[4] J. Hartigan. *Clustering Algorithms*. John Wiley & Sons, New York, 1975.

[5] D. J. Lehmann, G. Albuquerque, M. Eisemann, M. Magnor, and H. Theisel. Selecting coherent and relevant plots in large scatterplot matrices. *Comp. Graph. Forum*, 31(6):1895–1908, Sept. 2012.

[6] J. Schneidewind, M. Sips, and D. Keim. Pixnostics: Towards measuring the value of visualization. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 199–206, Baltimore, MD, 2006.

[7] J. Seo and B. Shneiderman. A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization*, 4(2):96–113, July 2005.

[8] L. Wilkinson, A. Anand, and R. Grossman. High-dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1363–1372, 2006.

[9] J. Yang, A. Patro, S. Huang, N. Mehta, M. O. Ward, and E. A. Rundensteiner. Value and relation display for interactive exploration of high dimensional datasets. In *Proceedings of the IEEE Symposium on Information Visualization*, INFOVIS '04, pages 73–80, Washington, DC, USA, 2004. IEEE Computer Society.