Exploring High Dimensional Data Through Locally Optimized Viewpoint Selection

Chufan Lai[1]* Xiaoru Yuan[1,2]†

1) Key Laboratory of Machine Perception (Ministry of Education), and School of EECS, Peking University
2) Beijing Engineering Technology Research Center of Virtual Simulation and Visualization, Peking University

ABSTRACT

Dimension reduced projection is usually obtained via a global optimization. It gives a good overview of the data, but cannot satisfy different users with different focuses. That's because any local data could be distorted and misleading due to projection errors. To address this problem, we propose an interactive visualization method, to customize projections for better local analysis. First, we allow users to define their point of interest (POI) data. Then we generate local projections to minimize distortions regarding the POI. Multipe optimizations are provided for different analyses. We also reveal relationships among different POIs, by comparing their local projections. At last, our method is proved effective via case studies with a real-world dataset.

Index Terms: Dimension-reduced projection, local analysis, high-dimensional data.

1 INTRODUCTION

Dimension reduced projection is widely used as an overview of the data. It approximates the original distribution in a low-dimensional space, using global optimization methods. But due to projection errors, distortions do exist in local areas. They are hard to ignore when analyzing local data. Users are often not aware of such errors, and could be misled by distorted local information. Even if distortions are revealed, there is no way for users to control them.

This problem draws much attention in previous researches. Efforts have been made to improve projections according to local data. A major branch seeks the best projection to separate different classes [3]. But such technique requires predefined classification, thus not suitable for interactive analysis. In more recent works, users are allowed to define subsets [4] or local relationships [2] in a projection. However, they either lose context of the whole dataset, or get too dependent on users' a prior knowledge.

In this work, we propose an interactive visualization technique, to improve projections for better local analysis. To be specific, we allow users to define their POI, and generate locally optimized projections to show different aspects of the POI. We call such projections 'viewpoints', meaning that each one is a specific angle to look at the data. Our method searches for good viewpoints, and manages the tour among them.

2 METHOD DESCRIPTION

To start with, we present PCA result as the initial viewpoint. With any existing viewpoint, our method consists of three steps. First, user chooses his POI in the projection. Then we provide two featured viewpoints, to support different local analyses concerning the POI. User can look for a new POI in the new viewpoint, and go back to the first step. At last, results can be stored and retrieved in the viewpoint map, which is an overview of all viewpoints.

2.1 Choose a POI

Due to distortions, local relationships in a projection can be highly unfaithful. To help users choose theirs POIs more wisely, we need to inform them of projection errors. When user hovers on a point, we suppose he is interested in that data. We adjust colors of all other points accordingly, to indicate their real distances to the hovered data. Closer points get higher saturation and vice versa. Figure 1 (b) shows a case where local distortions are found using the above technique.

Knowing the more accurate data relationships, the user can be more certain when choosing POIs. He may focus on either a point or a group of points. The two cases are essentially different in granularity, and thus should be handled separately.

2.2 POI Enhanced Viewpoints

Given the POI, we need to know what information is of concern in the analysis. It decides what kind of viewpoints we should recommend.

2.2.1 POI Point

When the POI is a single datum, users are expressing 'data like this one'. On one hand, they want to find similar data in the projection. On the other hand, they need to know in which attributes the data are similar. Such tasks are usually difficult due to distorted distances. We define the following target function regarding the POI:

$$f(\mathbf{x}_P) = \sum_{i \neq P} w_i \left\| \mathbf{x}_i - \mathbf{x}_P \right\|_2$$

Here, $\mathbf{x}_P$ and $\mathbf{x}_i$ are data in the projection. They denote the POI and the other data respectively. w_i is the weight of each datum. New viewpoints are gained by optimizing such a target function.

Considering the above tasks, we assign higher weights to POI's neighbors in the original space. When the target function is maximized, local structure around POI is best preserved. We call such projection the **'Point-Maximize Viewpoint'**. Users can now find the truly similar data around the POI (see Figure 1 (c)). When the target is minimized, the result is called the **'Point-Minimize Viewpoint'**. By checking dimensions of the projection, users can know in which aspects the POI neihgbours are most similar.

2.2.2 POI Group

When the POI is a group of data, relationships within the group is of special interest. Users may concern about the similarity and dissimilarity between group members. It's also necessary to explain such relationships in attributes. We define the following target function, only considering distances within the group:

$$f(G_P) = \sum_{i,j \in G_P} \left\| \mathbf{x}_i - \mathbf{x}_j \right\|_2$$

Here, G_P represents the POI group. When the function is maximized, dissimilarity is enhanced in the projection, showing

*e-mail: chufan.lai@pku.edu.cn
†e-mail: xiaoru.yuan@pku.edu.cn

the inner-group diversities. We call it the **'Group-Maximize Viewpoint'**. When the function is minimized, similarity is enhanced. All group members will be drawn together. We call it the **'Group-Minimize Viewpoint'**. By checking dimension components, users can learn that, in which aspects the group members are similar / dissimilar to each other.

2.3 The Viewpoint Map

After getting a good viewpoint, users may want to store it for further analysis. Also, it's important to compare different POIs by comparing their viewpoints. Therefore, we provide the viewpoint map, a projection of viewpoints. Those with similar projection directions are distributed closer. If two POIs have similar viewpoints, they are featured in a similar aspect.

3 Case Study

We apply our method to a real-world dataset to demonstrate its effectiveness. The dataset is known as the Auto-MPG data [1], which is widely used in multivariate analysis. It contains 398 samples with 8 numeric / categorical attributes: mpg, cylinders, displacement, horsepower, weight, acceleration, model year, and origin.

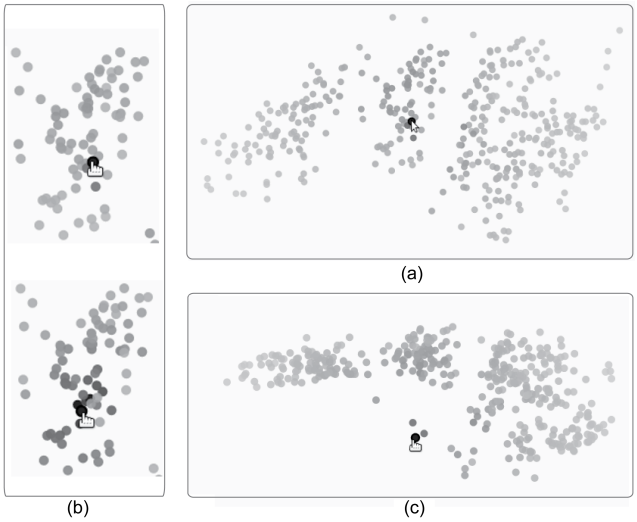

Figure 1: Discovery of an outlier. (a) The PCA result; (b) a central point is found inconsistent with its neighborhood; (c) a proper viewpoint shows that it's actually an outlier.

Figure 1 (a) shows the PCA result of the data. When hovering around in the projection, we discover that a strange datum is inconsistent with its neighborhood. When the point is hovered, its neighbors are not highlighted (see Figure 1 (b)). However, when one of the neighbors is hovered, most points in that area get affected. It suggests that this point doesn't belong to its neighborhood. To uncover the truth, we choose it as POI and generate the 'Point-Maximize Viewpoint'. The new viewpoint with less distortion is shown in Figure 1. It turns out the POI is an outlier, which confirms our previous guess. Moreover, it's easier to recognize real neighbors of the POI, which are also outliers. This case shows the effectiveness of distortion hints, as well as the point-based optimization.

Back in the PCA, data is roughly divided into three clusters. Assume that user is interested in the middle one. He wonders what are the common points shared in this cluster. So he brushes this part as a POI group (see Figure 2(a)), and applies the 'Group-Minimize Viewpoint'. In the new projection (see Figure 2 (b)), all data in the POI (darker points) are gathered more closely. It validates the effectiveness of our optimization. Moreover, the

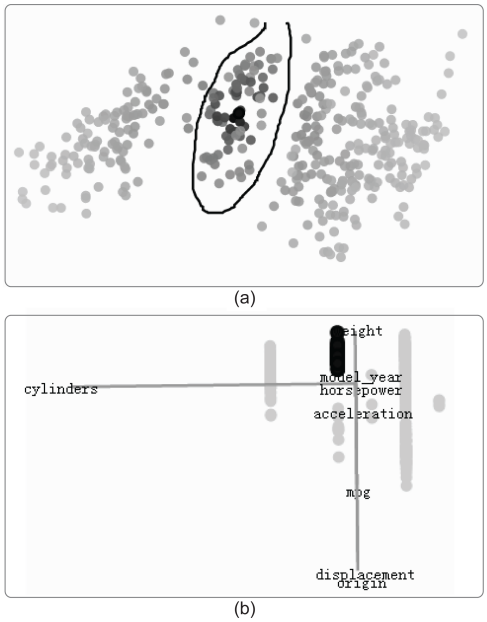

Figure 2: Analysis of a cluster. (a) User brushes a cluster as the POI group; (b) the optimized viewpoint shows strong patterns inherent to the cluster. Dimension components explain the clustering behavior.

whole distribution shows a stronger pattern. Some other clusters are discovered similar to the POI. Given the dimension components, we can see that three attributes: cylinders, displacement and origin, gain most of the weights. It helps users explain the grouping phenomenon.

4 Conclusion

In this work, we propose a method to generate locally-enhanced projections regarding any POI data. We belive that with a proper viewpoint, users can perceive more accurately, and learn more about local relationships. Such benifits are also confirmed in our case studies. At last, our method is not limited to high-dimensional data or linear projections. We plan to extend it to more general data and projections in the future.

5 Acknowledgement

This work is supported by NSFC No. 61170204, and partially funded by NSFC Key Project No. 61232012 and the National Program on Key Basic Research Project (973 Program) No. 2015CB352500. This work is also supported by PKU-Qihu Joint Data Visual Analytics Research Center.

References

[1] UCI Machine Learning Repository. `http://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data`.

[2] X. Hu, L. Bradel, D. Maiti, L. House, C. North, and S. Leman. Semantics of directly manipulating spatializations. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2052–2059, Dec 2013.

[3] M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan. Selecting good views of high-dimensional data using class consistency. In *Computer Graphics Forum*, volume 28, pages 831–838. Wiley Online Library, 2009.

[4] X. Yuan, D. Ren, Z. Wang, and C. Guo. Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2625–2633, Dec 2013.