Journal of Visual Languages and Computing 48 (2018) 144–156

Contents lists available at ScienceDirect

Journal of Visual Languages and Computing

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

View Points

Exploring high-dimensional data through locally enhanced projections

Chufan Lai[a], Ying Zhao[b], Xiaoru Yuan[*,a,c]

[a] Key Laboratory of Machine Perception (Ministry of Education), and School of EECS, Peking University, Beijing, 100871, PR China
[b] School of Information Science and Engineering, Central South University, Changsha, HB 410083, PR China
[c] Beijing Engineering Technology Research Center of Virtual Simulation and Visualization, Peking University, Beijing, 100871, P.R. China

ARTICLE INFO

Keywords:
Dimension-reduced projection
Local data analysis
High-dimensional data
Subspace analysis

ABSTRACT

Dimension reduced projections approximate the high-dimensional distribution by accommodating data in a low-dimensional space. They generate good overviews, but can hardly meet the needs of local relational/dimensional data analyses. On the one hand, layout distortions in linear projections largely harm the perception of local data relationships. On the other hand, non-linear projections seek to preserve local neighborhoods but at the expense of losing dimensional contexts. A sole projection is hardly enough for local analyses with different focuses and tasks. In this paper, we propose an interactive exploration scheme to help users customize a linear projection based on their point of interests (POIs) and analytic tasks. First, users specify their POI data interactively. Then regarding different tasks, various projections and subspaces are recommended to enhance certain features of the POI. Furthermore, users can save and compare multiple POIs and navigate their explorations with a POI map. Via case studies with real-world datasets, we demonstrate the effectiveness of our method to support high-dimensional local data analyses.

1. Introduction

Dimension-reduced projections are widely used for high-dimensional data analysis. They approximate distributions of high-dimensional data in low-dimensional spaces. Such approximations are often made as global ones that improve the overall mapping by striking a balance among all data. Two well-known examples are Principle Component Analysis (PCA) and Multidimensional Scaling (MDS). They generate good overviews of the data, but still unable to preserve all information without any loss [1]. Local distortion is an example of such loss, where inaccurate distance mapping may lead to unfaithful interpretations of data relationships [2,3].

Users are often not aware of the existence of distortions, and hence easily get misguided [4]. Even when distortions are informed [5,6], there are seldom interactive approaches for users to control them [7]. As a result, users often find it difficult to trust the projections [4,8]. On the other hand, users may wish to observe some local POI regions more precisely, while not so concerned about the other data. It inspires us to develop a dimension reduction scheme where users are able to decide which part of the projection is more precise and can be trusted.

Over the last few decades, different kinds of non-linear dimension reduction techniques have been developed to promote local data analyses [9–12]. Recent works further allow users to control the local mapping quality during a progressive rendering process [13]. Despite their abilities to preserve local structures, non-linear projections are not designed to visualize dimensional information. Cheng et al. [14] proposed to use interpolation and iso-contours for displaying attribute values. However, iso-contours may not be as simple and intuitive as axes in conveying dimensional semantics [15,16]. They are also prone to overlapping when the dimensionality increases. In this work, we choose to stay in the linear framework, since the linear projections provide intuitive dimensional semantics, are generally simple to use and interpret, and are also computationally efficient.

For the linear projections, Choo et al. proposed to preserve local structures by including supervised dimension reduction [17,18]. Along with similar works focusing on machine learning [19] and quality metrics [20,21], these approaches require the knowledge of data classification. It makes them unsuitable for general data explorations where no prior knowledge should be assumed. On the other hand, explorational methods allow users to manually adjust dimension weights of the projection [22–24]. However, the parameter search is often blind and time-consuming. It could be difficult to find a satisfying projection for a certain POI. Yuan et al. [25] proposed a framework where users are able to create new projections for data subsets. But the approach still requires a manual search of dimensional subspaces. In comparison, our approach generates projections and subspaces based on users' POIs and

* Corresponding author.
E-mail addresses: chufan.lai@pku.edu.cn (C. Lai), zhaoying@csu.edu.cn (Y. Zhao), xiaoru.yuan@pku.edu.cn (X. Yuan).

https://doi.org/10.1016/j.jvlc.2018.08.006
Received 21 July 2018; Accepted 11 August 2018
Available online 21 August 2018
1045-926X/

tasks. Users steer the projections by choosing their interests, without the need for any manual search.

In this paper, we propose an interactive scheme to help users customize a linear projection based on their POI data and analytic tasks. Specifically, users are able to specify a focus in the projection, which could be a single datum or a group of data. Then we offer multiple ways to alter the projection to enhance different features/aspects of the focus while maintaining the other data as contexts. Based on the locally enhanced projection, we further reveal dimensional subspaces that are most likely related to the features. It helps to interpret the features in the context of dimensions. In addition, we provide various means to support the data exploration at different stages. Users are assisted to discover, analyze, modify and compare different focuses. In summary, our contributions include:

- Given the user-defined POI data, we provide linear projections with enhanced POI features to support different kinds of local analytic tasks.
- Our method supports an interactive high-dimensional data exploration, where users are assisted to discover, analyze, modify and compare interesting pieces of local data.

The remainder of this paper is structured as follows. In the next section, we briefly review the related literature. Section 3 gives an overview of the proposed method. Then we elaborate each part of the method in detail in Section 4. Section 5 presents case studies to demonstrate the effectiveness of our method. In Section 6, we discuss weaknesses and potential improvements. At last, we end this paper with the conclusions.

2. Related work

Out method facilitates local data analysis in linear projections. We adopt the strategy of feature-driven projection pursuit [26,27], as opposed to dimension-driven methods [22–24]. We will briefly introduce the related works.

2.1. Data locality analysis in projections

Data locality has been extensively studied in high-dimensional data research. There are roughly two branches focusing on different aspects.

The major branch aims to improve global projections, focusing on preserving data localities. Many non-linear projections have been proposed for this purpose, such as Laplacian Eigenmaps (LE) [9], Locally Linear Embedding (LLE) [10], Local Tangent Space Alignment (LTSA) [11] and the T-distributed Stochastic Neighbor Embedding (t-SNE) [12]. These methods are fit for data lying on a low-dimensional manifold (e.g. face images of the same person), but the semantics of dimensions are lost. Cheng et al. [14] proposed to visualize attribute values using isocontours. But given their discrete natures, contours are not as intuitive as axes in conveying dimensional information [15,16]. They are also prone to overlapping when there are multiple layers. In comparison, our method keeps all projections in the linear framework. It helps users intuitively perceive and interpret the dimensional semantics of data relationships.

Another branch aims to reveal distortions in a projection. Martins et al. [28] examined distortions in different types of projections. They used color mapping to indicate distortion levels, and searched for real neighbors using automatic algorithms. Liu et al. took a step [6] further by analyzing data structures based on distortions. But none of them provide means to correct the distorted layout. Stahnke et al. [7] proposed a simple correction by directly mapping distances to the POI. Effective as it is, the approach loses dimensional contexts and is not suitable for situations where the POI is a group of data.

2.2. Projection assisted data exploration

Dimension reduced projections are often used to explore high-dimensional data. They are intuitive overviews, but hard to be changed interactively. Jeong et al. [22] proposed to change a projection by updating dimension weights in the PCA algorithm. Nam et al. [23] further enable users to freely decide the dimension components of a projection. Beyond parameter tuning, Lehmann et al. [24] proposed a more intuitive interaction, with which users can alter the dimension axes while maintaining an orthogonal mapping. These methods are indeed effective in updating a projection, but users need to go through a trial-and-error process to learn about the unpredictable effects of parameter changes. In comparison, our method helps users choose local POIs and their enhanced features, rather than dimension weights. Users are able to directly decide and predict the outcomes.

In a projection assisted exploration, subspace clusters are often provided beforehand [23,29,30]. In other methods [29,31,32], users can further participate in the clustering process. But in either way, users don't fully understand the given clusters or subspaces. It's hard for them to modify the results, let alone discovering more hidden clusters. Yuan et al. [25] proposed a hierarchical subspace exploration, which allows users to analyze a local subset in different subspaces. The approach helps to discover hidden clusters, but it doesn't provide any guidance for subspace selection.

2.3. Feature driven projection selection

Projection pursuit [26,27] is a well-known technique for finding interesting projections. It generates a series of projections to optimize a certain index. Gleicher et al. [19] used machine learning to train compositive dimensions for classification. Choo et al. [18] made the process interactive by involving users in a semi-supervised Linear Discriminant Analysis (LDA) process. In both works, user-defined classes are imported as the pursuit index. Apart from class labels, user-defined layouts can also function as the pursuit indices [33–35]. However, these methods require prior knowledge of the data, which cannot be assumed in a data exploration.

The rank-by-feature framework [36] is a variant of projection pursuit. It ranks existing projections according to feature strengths. Various kinds of metrics [37] are defined to measure different features, including class separation [20,21], clustering/outliers [38,39], and more complex topological properties [40]. They are helpful for analyzing a large group of scatterplots [41,42]. But most of them are result-oriented and computationally expensive, and thus unsuitable to guide the generation of projections. Otherwise, the time spent to find and score a projection will be unbearable in an interactive exploration. In this work, we only consider simple metrics when pursuing projections with desired features. The simple criteria are not only more efficient, but also easier to interpret.

3. Overview

In this work, we aim to facilitate local data analysis in linear projections. We propose an interactive and exploratory scheme to help users discover, analyze, modify, and compare different POI local data. To be specific, our method supports a four-step data exploration (Fig. 1):

Step. 1: **Focus Search**: First, we present a global projection as an overview (Fig. 1(a)) of the data. Users can choose any data subset in the layout and name it as the POI, which is also called a **focus**. We define two types of focuses, i.e. the focus point and the focus group, regarding whether the POI includes multiple samples. We also make recommendations for both types. Users can simply follow our suggestions if they don't know what to choose.