Journal of Visual Languages and Computing 48 (2018) 52–60

Contents lists available at ScienceDirect

Journal of Visual Languages and Computing

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

Exploring linear projections for revealing clusters, outliers, and trends in subsets of multi-dimensional datasets

Jiazhi Xia[a], Le Gao[a], Kezhi Kong[b], Ying Zhao[*,a], Yi Chen[c], Xiaoyan Kui[a], Yixiong Liang[a]

[a] School of Information Science and Engineering, Central South University, Changsha, China
[b] State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China
[c] Beijing Key Laboratory of Big Data Technology for Food Safety, Beijing Technology and Business University, Beijing, China

ARTICLE INFO

Keywords:
Multi-dimensional data
Projection
Visual exploring

ABSTRACT

Identifying patterns in 2D linear projections is important in understanding multi-dimensional datasets. However, local patterns, which are composed of partial data points, are usually obscured by noises and missed in traditional quality measure approaches that measure the whole dataset. In this paper, we propose an interactive interface to explore 2D linear projections with visual patterns on subsets. First, we propose a voting-based algorithm to recommend optimal projection, in which the identified pattern looks the most salient. Specifically, we propose three kinds of point-wise quality metrics of 2D linear projections for outliers, clusterings, and trends, respectively. For each sampled projection, we measure its importance by accumulating the metrics of selected points. The projection with the highest importance is recommended. Second, we design an exploring interface with a scatterplot, a projection trail map, and a control panel. Our interface allows users to explore projections by specifying interested data subsets. At last, we employ three datasets and demonstrate the effectiveness of our approach through three case studies of exploring clusters, outliers, and trends.

1. Introduction

Multi-dimensional data visualization plays an important role in data exploring and understanding. Among a variety of visualization approaches, 2D linear projection remains the most popular method to provide insights into structures and patterns in datasets [1]. Specifically, users are interested in the visual patterns of clusters, outliers, and trends in linear projections [2,3]. However, it is considered to be a fundamental challenge to identify interesting projections from the numerous possible projections [1].

To resolve this issue, several approaches have been proposed to provide a small set of representative projections. First, quality measures are adopted to rank possible projections [4]. Quality measures of clusters(e.g. Linear Discriminant Analysis [5]), trends (e.g. the Pearson correlation coefficient), and outliers (e.g. statistics analysis [4]) are widely studied. Specifically, the scagnostics [6] comprises nine measurements describing the patterns of points in projections, including outliers, shape, trend, and density (e.g. clumpy). Second, dissimilarities among projections are measured to reduce the redundant in the recommendation set [7]. Alternatively, Liu et al. [1] look for local maximum projections to provide a representative set.

However, most existing quality measures are defined in projections of the whole dataset. Real-world datasets often contain multiple clusters and noises. Local patterns that exist in a subset can be obscured by other components or noises. For instance, it is highly improbable to present patterns of clusters that lie in different subspaces in a single projection. Therefore, it is challenging to provide insight into local patterns based on global quality measures.

Let us consider a typical exploratory analysis scenario. When users explore projections for interested patterns, the exploring process often contains three stages. First, users look around the projection space until a global or local pattern is observed. Because the exploring space is large and the dataset is usually complicated, it is non-trivial to achieve a projection with the clear pattern in this stage. More probably, users observe a noised pattern, such as a set of points which are densely gathered and mixed with sparsely distributed points. Second, this observation yields a hypothesis of the existence of the local pattern. Specifically, the hypothesis is composed of the pattern type (e.g. cluster, trend, or outlier) and the subset of points that form the pattern. Third, this hypothesis motivates consequent exploration operations to verify it. The loop of looking around, suggesting a hypothesis, and verifying the hypothesis is performed iteratively in the exploring process.

* Corresponding author.
E-mail addresses: xiajiazhi@csu.edu.cn (J. Xia), csugaole@csu.edu.cn (L. Gao), durantkong@zju.edu.cn (K. Kong), zhaoying@csu.edu.cn (Y. Zhao), chenyi@th.btbu.edu.cn (Y. Chen), xykui@csu.edu.cn (X. Kui), yxliang@csu.edu.cn (Y. Liang).

https://doi.org/10.1016/j.jvlc.2018.08.003
Received 20 July 2018; Accepted 6 August 2018
Available online 09 August 2018
1045-926X/

However, as illustrated above, global quality measures cannot support the analysis of local patterns. Furthermore, the hypothesis is yielded dynamically in the exploring stage. Therefore, a pre-identification of all possible patterns is usually impractical.

In this paper, we propose a visual interface to explore linear projections and reveal possible patterns of clusters, outliers, and trends. Our interface contains two linked views and a control panel. The projection view presents the working projection and the weights of salient dimensions. It supports a set of exploring operations including identifying points, adjusting weights of dimensions, and jumping to a dissimilar projection. The projection trail map presents the identified salient projections. The control panel allows users to set the parameters of the system. Our critical idea is to propose a voting-based projection recommendation algorithm. Similar to Liu et al. [1], we perform a sampling in the space of all possible 2D projections. Other than the quality measure for a projection, we define a set of point-wise quality measures to evaluate the contribution of a single point to a local pattern in a projection. When users identify a possible pattern, the specified points vote for each projection. The stronger pattern the specified points show in a projection, the higher voting value the projection has. Therefore, we recommend the projection with the highest voting value as the target projection and transform to it through a smooth animation. Users can verify their hypothesis in the recommended projection.

In summary, the main contributions of this paper include:

- We introduce a set of interactions in a sense of indicating local pattern, including sketching, lasso, and point selection, for projection exploration;
- We propose a set of point-wise quality measures to evaluate the contribution of a point to a local pattern, and consequently enable a voting based projection recommendation framework to find the most salient projection;
- We design a visual interface that facilitates exploring patterns of clusters, outliers, and trends in multi-dimensional datasets.

The rest of this paper is organized as follows. Section 2 reviews the related work. We present our visual system design in Section 3 and illustrate our voting-based projection recommendation algorithm in Section 4. In Section 5, we present three case studies. We discuss our work in Section 6 and conclude this paper in Section 7.

2. Related work

A large number of techniques are proposed to pursue one or a small set of informative projections that reveal important features. In this section, we divide related work into two categories including automatic and interactive approaches.

2.1. Automatic approaches

The automatic approaches contain three types, including axis-aligned projections, dimensionality reduction, and quality measures.

As an axis-aligned approach, Scatterplots Matrix (SPLOM) [8] decomposes the high-dimensional space into a matrix of axis-aligned 2D projections. While the number of projections remains too large to visualize, users find it hard to get all important features in axis-aligned projections.

Another type of approaches is linear dimensionality reduction. Most famous techniques include Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) [9], and a linear version of Multi-dimensional Scaling (MDS) [10]. Linear dimensionality reduction follows an optimization scheme and is usually solved using an eigen-decomposition framework. The target 2D projection is often spanned by the first two decomposed dimensions. For instance, PCA tries to preserve the maximum data variances in the target projection. However, there are few dimensionality reduction methods designed for pursuing specific patterns, such as clusters, outliers, and trends. Especially, LDA pursues a subspace in which the labeled classes are best separated. For unlabeled data, Wang et al. [5] adopted an optimization algorithm which combines LDA and clustering to separate clusters in the target projection. A more comprehensive discussion on the connection between dimensionality reduction and clustering is given by Wenskovitch et al. [11].

A large number of quality measures have been proposed to measure projections in different aspects [12,13]. Tukey proposed scagnostics to identify interesting axis-aligned scatterplots. Wilkinson et al. [6] extended this idea as graph-theoretic scagnostics and proposed nine quality measures that capture patterns including outliers, shape, trend, and density. The rank-by-feature framework [4] adopts statistic characteristics as ranking criteria for projections. Projection pursuit [14] defined the interestingness of a projection as its deviation from a normal distribution. Quality measures enable a set of practical frameworks for selecting informative projections. Liu et al. [1] used quality measures to rank sampled projections. Lehmann et al. [7] used the dissimilarity measure as the objective function to find an optimal set of projections. However, the quality measures mentioned above are based on the whole dataset. While we are interested in local patterns, it is non-trivial to apply these quality measures to our application. Similar to our approach, Liu et al. [15] introduced a set of point-wise measures. While their measures are derived from the objective functions of dimensionality reduction techniques, our measures represent the point-wise contribution to specific patterns.

2.2. Interactive exploration of projections

Besides automatic approaches, there is a strong need to interactively explore the projection space, because interactive visual tools can bring users into the loop [16–19]. Nam et al. [20] decomposed the exploration process into five major tasks. Inspired by their categorization, we divide the key techniques of projection exploration into four categories: (1) Traveling along a route which is specified by setting the source projection and the target projection; (2) Manipulating the projection by interactively adjusting parameters; (3) Building a sight map which arranges interested projections; (4) Computing the target projection.

Traveling techniques try to provide a smooth tour from the source projection to the target projection. The Grand Tour [21], one of the early approaches, was proposed to explore multi-dimensional data by means of animation, which connects a source view with a target view. Similarly, Lehmann et al. [22] proposed Orthographic Star Coordinates and showed data tours following the definition of orthographic projection. Elmqvist et al. [23] proposed to travel among scatterplots in a SPLOM.

Manipulating techniques allow users to look around the current projection by adjusting the projection parameters interactively and intuitively. Star Coordinates [24] support interactive exploration of arbitrary 2D projections by handling the dimension anchor points. Clusters, trends, and outliers in multi-dimensional data can be visualized using Star Coordinates [2]. *TripAdvisor*$^{N-D}$ [20] proposes an N-D touchpad to adjust the parameters. However, users are required to keep track of the touchpad and the projection at the same time. In subspace voyager [16], an improvement of *TripAdvisor*$^{N-D}$, a trackball interface is used to navigate the projection and enhanced with star-coordinates-like manipulation. InterAxis [25] enables users to control the projection by manipulating data points. Xia et al. [26] proposed a subspace exploration interface, which allows users to manipulate subspaces by selecting dimensions. To explore subspaces, Yuan et al. [27] provided a dual space interface enabling selection of dimensions and data points.

Sight map is proposed to arrange interested projection, represent the relationships among them, and provide a mental map to users. The interested projections are pre-computed [28–30] or recorded in the trail [16,20,31]. Usually, projections are computed using PCA or MDS according to subspace distances. Alternatively, Liu et al. [1] formulated