Journal of Visual Languages and Computing 44 (2018) 28–38

Contents lists available at ScienceDirect

Journal of Visual Languages and Computing

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

Cluster aware Star Coordinates

Kang Feng[a], Yunhai Wang[a], Ying Zhao[b], Chi-Wing Fu[c], Zhanglin Cheng[d,*], Baoquan Chen[a]

[a] Shandong University, China
[b] Central South University, China
[c] Chinese University of Hong Kong, Hong Kong
[d] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China

ARTICLE INFO

Article history:
Received 1 September 2017
Revised 13 October 2017
Accepted 25 November 2017

Keywords:
Dimensionality reduction
Visual clustering
Star coordinates
High-dimensional data

ABSTRACT

Star coordinates is an important visualization tool for exploring high-dimensional data. By carefully manipulating the star-coordinate axes, users can obtain a good projection matrix to reveal the cluster structures in the high-dimensional data. However, finding a good projection matrix through axes manipulation is often a very tedious and trial-and-error process. This paper presents *cluster aware star coordinates plot*, which not only improves the efficiency of axes manipulation with higher cluster quality, but also enables users to learn the relations between cluster and data attributes. Based on the proposed approximated visual silhouette index, we introduce the silhouette index view, which interactively informs the user of the cluster quality of the projection. However, the user may still have no clue on how to manipulate the axes to improve the cluster quality. To resolve this issue, we propose a dimensionality reduction technique for visualization to progressively modify the projection matrix and improve the cluster results. Through this technique including a family of cluster-aware interactions, users can highlight important features of interest, such as points, clusters and dimensions, effectively investigate the change of cluster structures, and steer their relationship with the dimensions. In the end, we employ twelve high-dimensional data sets and demonstrate the effectiveness of our method through a series of experiments: comparison with state-of-the-art methods, interactive outlier detection, and exploration of cluster-dimension relationship.

1. Introduction

High-dimensional data has become very common in many application domains, such as information retrieval, computational biology, text mining, etc. To gain insight from these data, dimensionality reduction (DR) is often applied to reduce the data dimensions while maintaining the data features of interest (e.g., covariance and correlation between the data sets), so that we can plot and visualize the transformed data, e.g., using scatter plots [7,36]. If the data is unlabeled, we may also apply a clustering algorithm to generate the class labels, so that users can interactively explore the clusters and examine the data details.

However, treating DR as a black box to cluster exploration hinders the user to learn how individual dimension contributes to the final visualization results. Moreover, this irreversible scheme prevents us from incorporating user knowledge from the application domain and cluster exploration into the DR methods. This further restricts interactive visual cluster analysis of the high-dimensional data.

Rather than relying on machine analysis alone, star coordinates [21,22] help to incorporate user knowledge into the DR process by providing the user with a number of data transformation operations. These linear transformations enable us to plot the axes and visualize the transformed data in 2D/3D space. Through trial-and-error axes manipulation, the user can arrive at a projection matrix to better reveal the cluster structures of the high-dimensional data. However, finding a good projection matrix through axes manipulation is often a very tedious and trial-and-error process. We identified two reasons for this. First, the usual interfaces for axes manipulation are not constrained, say by the data features of interest. The lack of guidance forces the user into a trial-and-error mode of interaction, in which the axes are explored only by observing the changes in the results. Second, axis operation is a global transformation; axes manipulation cannot explicitly modify specific spatial relationship among the clusters. This can be frustrating if the user is interested in separating some spatially adjacent clusters that are of higher interest in the problem domain.

This paper presents the *cluster-aware star coordinates plot*, which not only improves the efficiency of axes manipulation with higher cluster quality, but also allows the user to learn the relationship between cluster and data attributes. To achieve this, we

* Corresponding author.
E-mail address: zl.cheng@siat.ac.cn (Z. Cheng).

https://doi.org/10.1016/j.jvlc.2017.11.003
1045-926X/

first adopt the silhouette index (SI) [33] to effectively evaluate cluster quality in star coordinates. By this, we construct the *SI view* in the GUI to inform users of how good the projection is in real-time. However, the user may still have no clue on the axes manipulation, which is trial-and-error and highly tedious, especially for data with large number of dimensions. Therefore, we propose a semi-automated approach to progressively modify the projection and improve the cluster quality by combining user controls and a lightweight optimization model. We achieve this by efficiently computing a set of next best positions (NBP) of the axes configuration using a *fast greedy model* to locally refine the results and letting the user iteratively select and modify the axes. Hence, we can incorporate user knowledge into the DR process and provide visual guidance to users, thus the user can focus on the global axes manipulation (e.g., which axis to pick and what clusters to explore) with their knowledge without needing to deal with tedious refinement. In addition, to reduce the high computational cost of SI for supporting the interaction, we propose an *approximated version of SI*, which can produce similar results at a reduced cost, from $O(n^2)$ to $O(nk)$, where n and k are the number of data points and clusters, respectively.

Besides the above dimensionality reduction technique, we develop a family of *cluster aware interactions* to help users investigate the cluster structures and steer their relationship with the data dimensions. First, the user can move axes to learn how each data attribute impacts the cluster structures. To provide visual hints to aid the manipulation, we visualize the estimated NBP by a *compass metaphor*. Second, the user can assign high importance values to points of interest by a simple brushing operation and visualize how the axes configuration changes. This facilitates the user to learn why some points are mis-classified, see experiment in Section 6. Lastly, the user can also assign high importance values to clusters of interest, and study which attributes can increase/decrease the separation between clusters. These importance values are integrated into the energy function in the optimization, which can be solved by methods such as conjugate gradient [32], but experimentally, we found that a simple solution that always picks the best axes to manipulate would produce good and efficient local refinement results, and achieve high performance to support the interactions.

In summary, the main contributions of this paper include:

- We introduce a dimensionality reduction technique for visualization, where the user can focus on the global axes manipulation with their domain knowledge, and rely on the interface to locally refine the cluster quality;
- We develop a family of cluster aware interactions for studying the relations among the cluster structures and data attributes (equivalently, the data dimensions);
- We propose a new approximated visual silhouette index to support fast computation of cluster quality and interactive star coordinate visualization of high-dimensional data.

2. Related work

2.1. Dimensionality reduction techniques

Dimensionality reduction (DR) methods can be categorized into linear and non-linear methods. Assuming that the data lives close to a low-dimensional linear subspace, linear DR methods project the data to a lower dimensional space by a linear transformation. There is an abundance of methods [10] that aims to preserve different data features of interest. Among them, principal component analysis (PCA) [19] and linear discriminant analysis (LDA) [19] are two most commonly used methods due to their relative simplicity and effectiveness in the computation. PCA maximizes the data variance captured by the low-dimensional projection, while LDA maximizes the separation of classes in labeled data. To combine their advantages, Choo et al. [8,9] propose a two-stage framework for the visualization of labeled data, where LDA is first used to obtain a cluster-preserved low-dimensional data, and PCA is then applied to further reduce the dimension to 2 for visualization. These linear methods are computationally efficient, but they often miss non-linear structures within the data. An exception is locality preserving projections (LPP), which aims to preserve the neighborhood structure of the data; however, it cannot maximize the cluster separation. All these methods can be used to initialize the projection matrix in our method, and we seek the one that maximizes the cluster separation in the 2D visualization space.

Multi-dimensional scaling (MDS) is another widely used DR approach, which attempts to preserve the dissimilarities between high-dimensional objects in a lower dimensional space. Since it involves an $O(n^2)$ optimization, many methods have been proposed [5,18] to accelerate the computation. To better preserve neighborhood structure, manifold methods uncover the intrinsic structure by building a model of manifold connectivity. Examples include Isomap [43], Locally linear embedding [34], Laplacian Eigenmaps [4], and many other variants. Unfortunately, these methods are computationally intensive and heavily rely on the constructed neighbourhood graph. By transforming distances of data points to probabilities, the recent proposed method t-SNE [46] and Barnes-Hut-SNE [45] can produce some visualizations with well-separated clusters of high-dimensional data. In this paper, we compared our method with Barnes-Hut-SNE and demonstrated that we can produce similar or even better results in less time.

To incorporate user knowledge into the construction of the projection, some user-driven projection techniques [31] have been proposed. Among them, partial linear multi-dimensional projection [30] and local affine multi-dimensional projection [20] both allow the user to position anchor points and then find a projection that meets the constraints. However, these methods cannot explain how the clusters are separated by the data attributes. Gleicher [14] proposes crafted projections, where the pursued projection attempts to explain the relationship between the data attributes and user-defined concepts. Kim et al. [23] provide observation-level interactions to maintain the interpretability of axes in a scatter plot, which presents the outputs of the DR techniques. Similar to these techniques, our method also supports user-assisted projection, but it is enabled by interactive user controls, where the user can define various importance by brushing and explore the relations between the cluster structures and data attributes.

2.2. Exploring projection quality

DR provides a means to explore structures hidden in high-dimensional data, but it produces inevitable projection errors. These errors can be measured with “stress” [24] of each point and visualized by coloring the corresponding Voronoi cell in the projection space [2]. By combining heat map and height map, Seifert et al. [38] present stress maps to display local stress values, which are used for artifact identification. Recently, Stahnke et al. [42] introduce a set of interaction techniques that integrates stress examination with data exploration together.

Besides stress, the projection error can also be inspected by other metrics. For example, Lespinats and Aupetit [27] visualize the distortion by measuring tears and false neighbourhoods. If the class label is available, the metrics of cluster separation can be used [41]. Sedlmair et al. [37] develop a taxonomy of visual cluster separation factors for projection quality evaluation. In this paper, we not only visualize the goodness of clusters in the DR results