Journal of Visual Languages and Computing 33 (2016) 3–12

Contents lists available at ScienceDirect

Journal of Visual Languages and Computing

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

View Points

Two axes re-ordering methods in parallel coordinates plots

Liang Fu Lu [a,b], Mao Lin Huang [c,b,*], Jinson Zhang [b]

[a] Mathematics Department, Tianjin University, Tianjin, PR China
[b] Faculty of Engineering and IT, University of Technology, Sydney, Australia
[c] School of Computer Software, Tianjin University, Tianjin, PR China

ARTICLE INFO

Article history:
Received 19 November 2015
Accepted 1 December 2015
Available online 17 December 2015

Keywords:
Multidimensional data visualization
Visual analytics
Parallel coordinates
Axes re-ordering
Singular value decomposition
Nonlinear correlation coefficient

ABSTRACT

Visualization and interaction of multidimensional data are challenges in visual data analytics, which requires optimized solutions to integrate the display, exploration and analytical reasoning of data into one visual pipeline for human-centered data analysis and interpretation. Even though it is considered to be one of the most popular techniques for visualization and analysis of multidimensional data, parallel coordinate visualization is also suffered from the visual clutter problem as well as the computational complexity problem, same as other visualization methods in which visual clutter occurs where the volume of data needs to be visualized to be increasing. One straightforward way to address these problems is to change the ordering of axis to reach the minimal number of visual clutters. However, the optimization of the ordering of axes is actually a NP-complete problem. In this paper, two axes re-ordering methods are proposed in parallel coordinates visualization: (1) a contribution-based method and (2) a similarity-based method.

The contribution-based re-ordering method is mainly based on the singular value decomposition (SVD) algorithm. It can not only provide users with the mathmetical theory for the selection of the first remarkable axis, but also help with visualizing detailed structure of the data according to the contribution of each data dimension. This approach reduces the computational complexity greatly in comparison with other re-ordering methods. A similarity-based re-ordering method is based on the combination of nonlinear correlation coefficient (NCC) and SVD algorithms. By using this approach, axes are re-ordered in line with the degree of similarities among them. It is much more rational, exact and systemic than other re-ordering methods, including those based on Pearson's correlation coefficient (PCC). Meanwhile, the paper also proposes a measurement of contribution rate of each dimension to reveal the property hidden in the dataset. At last, the rationale and effectiveness of these approaches are demonstrated through case studies. For example, the patterns of Smurf and Neptune attacks hidden in KDD 1999 dataset are visualized in parallel coordinates using contribution-based re-ordering method; NCC re-ordering method can enlarge the mean crossing angles and reduce the amount of polylines between the neighboring axes.

* Corresponding author at: Faculty of Engineering and IT, University of Technology, Sydney, Australia.
E-mail address: Mao.Huang@uts.edu.au (M.L. Huang).

http://dx.doi.org/10.1016/j.jvlc.2015.12.001
1045-926X/

1. Introduction

The rapid growth of data communication through the Internet and World Wide Web has led to vast amounts of information available online. In addition, business and government organizations create large amounts of data containing both structured and unstructured information which needs to be processed, analyzed, and visualized. Therefore, multi-dimensional data analysis is becoming a commonplace as the number of applications increases, such as statistical and demographic computation, digital libraries and so on. However, traditional visualization techniques for these data sets usually require dimensionality reduction or selection to generate the meaningful visual representations. Dimensionality reduction, as Sara Johansson et al. [1] pointed out, is always employed prior to visualization for dealing with the data with a large number of attributes. Currently, many dimensionality reduction methods are available to preserve the information inside the data as much as they can remove some less relevant data items or attributes from the original dataset. While dimension selection is mainly referred to dimension re-ordering which means the corresponding axes of the dimension in a parallel coordinate visualization can be positioned in accordance to some effective rules such as similarity of dimensions to achieve good visual structures and patterns. This paper focuses on the dimension re-ordering instead of dimension reduction to address the problems of visual clutter and computational complexity.

In 1998, Mihael Ankerst et al. [2] presented the method of using the similarity of dimensions to improve the quality of visualization of multidimensional data, that is using global and partial similarities for one or two-dimensional visualization methods. Pearson's Correlation Coefficient (PCC) is one of the most common methods used for measuring similarity between two dimensions. PCC can be used for dimension reduction, clutter reduction and clustering in data visualization. At the same time, it has also been proved that the PCC based re-ordering problem is a NP-complete problem. Therefore, many researchers applied heuristic algorithms to figure out an optimal order of axes (or dimensions) in the multi-dimensional data visualization.

This paper proposes two rational dimension re-ordering methods to support the visual analytics in parallel coordinates. And those two methods can also be easily applied to other visualization techniques.

Firstly, method to find out the contribution of each dimension in the dataset is developed on the basis of the Singular Value Decomposition (SVD). After the calculation of contribution rates, axes (or dimensions) are re-ordered and visualized as parallel coordinates from left to right according to the degree of their significances. Though the traditional heuristic algorithms can optimize the order of axes for one- or two-dimensional visualizations, most studies have not been done to a deeper investigation on how to determine the first dimension (the most significant dimension) in multi-dimensional data visualizations. The first dimension always attracts much more user's attention than the others. Therefore, the one with the highest contribution rate can be considered as the first dimension to simplify the traditional similarity-based re-ordering methods and the one to find out the optimal order of parallel axes in a short time period.

Secondly, Pearson's Correlation Coefficient method is applied into the further investigation on axes re-ordering. As a correlation metric between each pair of dimensions in the dataset, PCC is available for characterizing linear systems statistically. Inspired by PCC, a similarity-based re-ordering method is presented in parallel coordinates which is based on the combination of a Nonlinear Correlation Coefficient (NCC) and the SVD algorithms. NCC is sensitive to any relationship, not just the linear dependence [26]. It is more rational than the current PCC method in theory and it can improve the quality of multi-dimensional visualizations significantly in terms of effectiveness and exactness. In our experiments, the effectiveness of the new method can be proved by visualizing the patterns and enlarging the mean crossing angles for better visual representation.

The paper is organized as follows. The current situation and background of researching on similarity measure and dimension reordering in high-dimensional data visualization is introduced in Section 1. And next two dimension reordering approaches are stated in detail in Section 2. The experimental evaluation for our new ideas as well as the effectiveness of our methods in parallel coordinates visualization is further elaborated and proved in Section 3. Section 4, the last part of the paper, is designed to make the conclusions and look forward to future work.

2. Related work

An effective way to improve the quality of multi-dimensional visualizations is to re-order the dimension axes in parallel coordinates based on similarity of data attributes. In this section, paper begins to summarize the previous researches finished in the area of high-dimensional visualization.

Parallel coordinates [3,4], scatter plot matrix [5], table lens [6] and pixel-oriented display [7] etc. are well-known and accepted as visualization techniques for high-dimensional data sets.

Similarity measurement as one aspect of quality metrics in high-dimensional data visualization has been addressed in the past few years [1,8–11]. It is worth noting that Enrico Bertini et al. [8] systematically presented an overview of quality metrics in many visualization techniques through a literature review of nearly 20 papers and considered correlation between two or more dimensions to be the main characteristic of similarity measurement. Sara Johansson [1] introduced a weighted quality metrics to their task-dependent and user-controlled dimensionality reduction system, where small correlation values are ignored to reduce the dataset that preserves the important