Neurocomputing 150 (2015) 599–610

Contents lists available at ScienceDirect

Neurocomputing

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

Projection inspector: Assessment and synthesis of multidimensional projections

Paulo Pagliosa [a], Fernando V. Paulovich [b,*], Rosane Minghim [b], Haim Levkowitz [c], Luis Gustavo Nonato [b]

[a] Federal University of Mato Grosso do Sul, UFMS, Brazil
[b] University of São Paulo, USP, Brazil
[c] University of Massachusetts Lowell, UML, USA

ARTICLE INFO

Article history:
Received 21 November 2013
Received in revised form
16 April 2014
Accepted 9 July 2014
Available online 27 October 2014

Keywords:
Multidimensional projection
Quality metrics

ABSTRACT

As the number and complexity of visualization techniques have grown, it has become progressively more difficult to make a decision as to which technique to employ for any given situation or application. A particular case is that of multidimensional data visualization utilizing projections, which have gained much attention lately and are being utilized in a growing number of applications. With their popularity, many new variations of multidimensional projections have been proposed in the literature. Numerical evaluations are varied and are useful, but do not reflect visual properties of projections accurately. In this paper we present *Projection Inspector*, an approach that contributes to the problem of understanding the difference amongst projections. It is an interactive assessment method that allows a user to explore a "space" of known projection techniques and view their results, as well as to identify the differences between them. In addition, it generates "on-the-fly" new projection techniques via interpolations of existing techniques as the user explores the projection space. We present the theoretical foundations of the projection exploration space and an interactive tool that implements a view of this space. We demonstrate the approach with case studies that demonstrate the need for projection assessment and the value of combining projections into new, better suited, projection alternatives.

1. Introduction

Multidimensional projection (MP) methods have developed significantly in the last decade; they are now considered among the most relevant methods to handle, analyze, and visualize high-dimensional data. Such growing interest has lead to the development of many MP methods, which vary considerably in their mathematical foundations, optimization criteria, and computational complexity. One of the reasons for the proposal of so many distinct methods is that the effectiveness of MP methods in revealing valuable information contained in a data set depends strongly on the nature of the data. Moreover, quality metrics for MP methods differ considerably in the property they measure, making it harder for the user to choose what method fits their problem.

The problem of selecting an adequate multidimensional projection from a set of possibilities has been treated in the literature. However, the existing alternatives either rely on multiple views generated from a single MP method, such as Projection Pursuit [1] and ranked scatter plot matrices [2], or are not flexible enough to allow a free navigation throughout the possibilities, as is the case of Stress Maps [3] and CheckVis [4]. When faced with deciding what method to adopt for his or her data set and task, the analyst is often faced with various views, various numbers and in the end is not sure what alternative is offering a proper compromise in each case. In that respect, existing solutions are not quite satisfactory and there is a clear need for more flexible methods to navigate through feasible MP methods applied to a particular data set, while still examining multiple quality metrics.

We propose a novel method to interact with multiple projections and quality metrics that supports the user towards selecting the most appropriate projection in terms of both quality metric and layout organization. Our method, named Projection Inspector (ProjInspector), allows not only to easily navigate through a set of different projections but also to combine them when the best solution for a data set is not given by any single projection. ProjInspector does that by generating MPs from the combination of basic MP methods and enabling visual inspection of their layouts as well as of their quality metrics. Smoothing transition between distintic layout combinations is ensured by using control points to register basic layouts, thus avoiding drastic jumps during user navigation.

* Corresponding author.
E-mail address: paulovic@icmc.usp.br (F.V. Paulovich).

http://dx.doi.org/10.1016/j.neucom.2014.07.072
0925-2312/

We demonstrate the usefulness of ProjInspector with a set of experiments showing how the accuracy of projections change depending on the adopted metric. We also show that sometimes the quality compromise the user is looking for to express his or her data lies in a combination of projections instead of in any particular layout. Additionally, we illustrate how ProjInspector can also be used to analyze how parameters affect the quality of a projection, thus enabling the interactive selection of the best set of parameters for an MP method when dealing with a particular data set. These goals are achieved by interpolating between basic projections and giving the user support to select a method or a combination of methods for their case.

In summary, the main contributions of this paper are:

- A novel method for analyzing and combining of multidimensional projection methods, enabling the construction of families of MP layouts.
- An interactive tool to navigate easily through a set of MP layouts, which enables visual inspection of the quality of each particular layout.
- A set of experiments showing that the quality of a projection changes drastically according to the metric and the data set, thus demonstrating the need for a tool like ProjInspector.

Besides the main contributions mentioned above, we also propose a variant of currently available neighborhood preservation metrics, called *Smooth Neighborhood Preservation* (SNP), which takes into account the number of neighbors preserved in the projection as well as the distance that misplaced points are from their correct position.

2. Related work

The need and desire to compare and evaluate different multidimensional projections, both qualitatively and quantitatively, is not new. Published research has taken two distinct directions: correlating multidimensional projection quality metrics with human perception, and offering multiple projections to assist users in the analysis of high-dimensional data. In order to better contextualize our approach we focus the related work discussion on the latter class of methods, that is, methods that make use of multiple projections to analyze high-dimensional data. Readers interested in the former class of methods can refer to [5–9].

We group techniques that rely on multiple projections into three main categories: multi-projection tour techniques, projection boards, and multi-projections with distortion analysis.

Multi-projection tour techniques aim at enabling interactive and animation mechanisms to generate a set of projections that will assist users to "navigate" through high-dimensional data. As early as 1985 (even before the coining of the term "visualization") Asimov [10], introduced The Grand Tour, "a method for viewing multivariate statistical data via orthogonal projections onto a sequence of two-dimensional subspaces." However, the Grand Tour was limited to attribute driven scatter plots. As data dimensionality grows, the computational demands become prohibitive and the possible combination of axes troublesome. Dhillon et al. [11] introduced "class tours", which are sequences of two-dimensional, class-preserving projections of multidimensional data that are displayed in a rapid and smooth sequence. Class tours' purpose is to enable a "view" of higher-dimensional subspaces. In addition, class-similarity graphs that are overlaid on the projections provide a "skeleton of the data", guide the user through the projections, and help estimate distance relationships in the original high-dimensional space. The authors report to have a "mechanism" that is theoretically able to view inter-class relations of any subset q of k classes. In reality, however, as q grows, classes' distinction blur. Further, the method is not interactive: the tour is conducted through static photographs, which are pre-processed. Rolling-the-Dice [12] relies on a cube metaphor to interactively smooth out the transition between scatter plots during visualization. Dimension reordering and a sculpting mechanism is used to further improve the high-dimensional data exploration. One of the main issues with tour-based techniques is that quality metrics are not directly used to assist users during the data analysis. Moreover, those techniques strongly rely on attribute scatter plots, and cannot be easily extended to operate on other types of multidimensional projections methods.

Projection board methods differ from tour-based techniques in that they generate static boards where projections are arranged according to ranking criteria. Projection Pursuit [13] and its variants [14,1], for instance, generate a family of scatter plots ranked according to their concentration of points into clusters while preserving the separability of those clusters. Rank-by-Feature [2] is an interactive framework that allows users to select interesting dimensions according to distinct rank criteria, producing a set of scatter plots from user selections. Tatu et al. [15] automate the ranking process to generate scatter plots (and Parallel Coordinates) of classified and unclassified data according to data correlation and cluster separation. Their goal was to aid and potentially speed up the visual exploration process for different visualization techniques. Sips et al. [16] proposed to select "good views" of high-dimensional data by utilizing what they refer to as "class consistency." They measure class consistency by the distance of class members from the class's center of gravity and by the entropies of the spatial distributions of classes. In addition, they asked users to choose "good views," and report that class consistency demonstrated good precision and recall. They have evaluated their consistency measures using various data sets, and concluded that the measures appear to be efficient and robust. Schreck et al. [17] propose the use of a quantitative quality measures to filter out scatter plots so as to highlight the ones with higher precision. They also present a scheme that allows to locally analyze the quality of distinct projection methods as well as the quality of clusters. Lehmann et al. [18] present an interactive approach for generating ranked scatter plot matrices, where the elements are organized according to a reordering technique supported by quality measures. Lehmann's approach bears a set properties that are not present simultaneously in any other scatter plot-based method. Wang et al. [19] proposed a less interactive approach, a mechanism to first cluster dimensions based on their similarity and then filter out dimensions according to an importance criterion. A scatter plot matrix containing only the "important" dimensions is generated in the end of the process. A main issue with Wang's approach is the loss of context when navigating throughout the levels of the hierarchy. Similar to tour-based methods, projection board techniques strongly rely on attribute scatter plots. No significant research has attempted to extend those approaches to general multidimensional projections.

Multi-projection with distortion analysis techniques enrich projection layouts with mechanisms that enable the visualization of distortions during the dimensionality reduction process. Aupetit [20] proposes the use of colored Voronoi cells to visualize quality measures defined on projected points, segments connecting projected points, and triangles formed by projected points. These techniques allow visual identification of regions where the neighborhood of each point stretches or compresses. Colored Voronoi cells are also used by CheckViz [4] to visualize false neighbors introduced by the projection mechanism, allowing to compare the distortion introduced by different projection methods. ProxiLens [21] provides an interactive scheme to highlight and filter out false neighbors when visualizing projected data, making it easier to analyze the "true" neighborhood of the data. Stress Maps [3] make use of the landscape metaphor to visualize local stress