Journal of Visual Languages and Computing 33 (2016) 13–23

Contents lists available at ScienceDirect

Journal of Visual Languages and Computing

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

View Points

Enabling decision trend analysis with interactive scatter plot matrices visualization

Wen Bo Wang [a], Mao Lin Huang [a,b], Quang Vinh Nguyen [d], Weidong Huang [e], Kang Zhang [b,c], Tze-Haw Huang [a]

[a] School of Software, University of Technology, Sydney, Australia
[b] School of Computer Software, Tianjin University, Tianjin, China
[c] Department of Computer Science, University of Texas at Dallas, Richardson, TX, USA
[d] MARCS Institute and School of Computing, Engineering and Mathematics, University of Western Sydney, Australia
[e] School of Engineering and ICT University of Tasmania, Australia

ARTICLE INFO

Article history:
Received 19 November 2015
Accepted 24 November 2015
Available online 2 December 2015

Keywords:
Rough set theory
Dimensionality reduction
Scatter plot matrices
Parrallel coordinate geomery
Visual data analytics
Visual decision making

ABSTRACT

This paper presents a new interactive scatter plot visualization for multi-dimensional data analysis. We apply Rough Set Theory (RST) to reduce the visual complexity through dimensionality reduction. We use an innovative point-to-region mouse click concept to enable direct interactions with scatter points that are theoretically impossible. To show the decision trend we use a virtual *Z* dimension to display a set of linear flows showing approximation of the decision trend. We conducted case studies to demonstrate the effectiveness and usefulness of our new technique for analyzing the property of three popular data sets including wine quality, wages and cars. The paper also includes a pilot usability study to evaluate parallel coordinate visualization with scatter plot matrices visualization with RST results.

1. Introduction

Multi-dimensional data exploration presents a great challenge to information visualization because features of data are inherently sparse in high dimensional data and the over-plotting of visual display makes it even difficult to observe any useful patterns. Visualization methods for large dimensional data are not usually effective due to density of high dimensions and the limitation of screen display. Interactive zooming could be used to provide an aid for exploring and reducing the number of dimensions. However, the interaction is still limited where some of the contextual information could be lost during the navigation.

The efficiency of knowledge discovery tends to decline while the processing cost of information interpretation tends to increase because some are noisy data and not necessary all the dimensions need to be analyzed. This phenomenon is also known as *curse of dimensionality* which was first apparently coined by Bellman [1] to describe the problem that data samples will grow exponentially according to the changes of the number of dimensions because of the necessity of fitting a multivariate function for a given degree of accuracy.

Dimensionality reduction is important in many application domains to be facilitated with classification, visualization of dealing with the complexity of multi-dimensional data. It reduces the intrinsic dimensionality of the data in order to cut down the cost of time and space complexities required for subsequent computation and analytic task. The terms *variable*, *feature* and *attribute* are commonly quoted in various research fields hence we would use them interchangeably throughout the paper.

Dimensionality reduction can be divided into feature selection and feature extraction. Feature selection is

E-mail address: mao.huang@uts.edu.au (M.L. Huang).

http://dx.doi.org/10.1016/j.jvlc.2015.11.002
1045-926X/

mainly to select a subset of the original variables according to selection principals. In the supervised method, the general criteria requires user to guide the selection process through choosing weighted quality metrics, therefore the selection rule would prefer the attributes weighted above the threshold. However in this case, user's expertize about quantization would have a great influence on the effectiveness of variable selection as quantization is typically not a trivial task. More importantly, empirical studies are the fundamental basis of applying quantization; hence the method may work well on this data set but might completely fail on another. On the other hand, feature extraction is a typically unsupervised technique with minimal consideration about user factors. The absence of user guidance raises the challenge of information interpretation if the result is unintuitive or not expected by the user that is often criticized as information loss. Most techniques developed in the past are projection based, implying that phenomena of interest higher than second order could not be discovered. Strictly speaking, projection means orthogonal. The oversimplified pattern is not adequate to support interactive data exploration that requires iterative interaction through visualization for the adjustment of input vectors to increase the accuracy of analytical results for decision trend analysis. Multi-dimensional data exploration via dimensionality reduction is really a user centric task in information visualization. Most dimensional reduction methods do not provide multiple results and make no assumption with the consideration of user's concern. Ideally, an effective method should only require the user to guide the procedures of dimensionality reduction, in terms of specifying a most concerned attribute and adjusting the values of input vectors subjectively.

In our previous works, we integrated Rough Set Theory (RST) with parallel coordinates [2] and scatter plot [33,34] for interactive feature selection. RST is a mathematical approach to data vagueness and uncertainty, which can be considered as discovering facts from complicated data through dimension reduction with a given dimension known as decision specified by the user. In this paper, we further extend our prior work with additional contributions described as follows:

- A feature ranking method on the result to guide the user for multi-dimensional data analysis.
- Interactive data exploration support in scatter plot matrices for class data.
- Enhanced scatter plot matrices for decision trend analysis.
- Provide more case studies to illustrate the visualization on different data sets.
- Carry out a pilot usability study on the visualizations.

2. Related works

There are several techniques for visualizing multi-dimensional data, such as Parallel Coordinate [3], Start Plots, Scatter plot Matrix [4], Mosaic Plots, Heat Map, Glyphs and Icons. Among them, Parallel Coordinate and Scatter plot Matrix are considerably popular techniques for large scale data sets. Theoretically, they are capable to visualize the data with unlimited number of dimensions nevertheless their visual efficiencies tend to decline when number of dimensions grows.

Some developments addressed the problem by visual transformation. Guo et al. [5] and Artero et al. [6] used clustering to highlight the patterns of homogenous data in parallel coordinate. Peng et al. [7] applied dimension reordering to rearrange the dimension axes based on visual neighboring similarity for clutter reduction. However, using visual transformation to enhance the visual structure still left data in high dimensional space with sparse features. Nguyen et al. [35] presented a multi-dimensional data visualization system based on scatter plot with flexible axis and attribute mapping. The tool also provided interaction, filtering, zooming and dynamically control to the visualization. Although these techniques are quite effective to visualize small numbers of dimensions, dealing with high numbers of dimensions remains a challenge.

The widely accepted dimensionality reduction methods are Principal Component Analysis (PCA) [8], Multi-dimensional Scaling (MDS) [9] and Self-Organizing Map (SOM) [10]. PCA is a linear transformation method that projects the original data onto a much smaller set without ordinarily result. The selection principles – are typically interested in dimensions with largest eigenvalues, known as principal components because they explain the majority of variability. The low dimensional view that represents the high dimensional dataset is formed by rotating the principal components along the linear directions of maximum variability. MDS aims to place the data points that the pairwise distances are preserved as well as possible. SOM is an unsupervised learning algorithm based on neural network model by reducing the dimensions to low-dimensional (typically 2D) layer of neurons. Locally Linear Embedding (LLE) [11] is another popular unsupervised learning technique that computes nearest neighborhood of each dimension to obtain the low dimensional embedding of high dimensional data. One common drawback of these methods is that they project the dataset into extremely low dimensions that could oversimplify patterns. For projecting an information correlated dataset i.e. survey dataset, into 2D space is usually meaningless for human centric knowledge discovery.

Projection Pursuit (PP) [12] is a type of statistical technique for the pursuit the choices of possible projections in multi-dimensional data that can reveal the most details about the structure defined by a projection index. The pursuit of the possible projections globally involves non trivial computational intensive task [13]. XGobi [14] is a visualization system that integrated PP for viewing high dimensional data. The choices of possible relevance are the commonality between our work and PP. The main problem of PP is the difficulty to quantize the value of projection index because it is possible to present spurious interesting structures with an inappropriate projection index.

Several Visual Dimensionality Reduction (VDR) methods have been proposed by taking advantages of information visualization at different stages. Yang proposed Visual Hierarchical Dimensionality Reduction (VHDR) [15] method by visually grouping dimensions into a hierarchy