Improved Visual Correlation Analysis for Multidimensional Data

Yi Zhang, Teng Liu, Kefei Li, Jawan Zhang

PII: S1045-926X(16)30171-9
DOI: 10.1016/j.jvlc.2017.03.005
Reference: YJVLC 777

To appear in: *Journal of Visual Languages and Computing*

Received date: 18 September 2016
Revised date: 8 February 2017
Accepted date: 23 March 2017

Please cite this article as: Yi Zhang, Teng Liu, Kefei Li, Jawan Zhang, Improved Visual Correlation Analysis for Multidimensional Data, *Journal of Visual Languages and Computing* (2017), doi: 10.1016/j.jvlc.2017.03.005

Improved Visual Correlation Analysis for Multidimensional Data☆

Yi Zhang, Teng Liu*, Kefei Li, Jawan Zhang

Tianjin University, China

Abstract

With the era of data explosion coming, multidimensional visualization, as one of the most helpful data analysis technologies, is more frequently applied to the tasks of multidimensional data analysis.Correlation analysis is an efficient technique to reveal the complex relationships existing among the dimensions in multidimensional data.However,for the multidimensional data with complex dimension features,traditional correlation analysis methods are inaccurate and limited.In this paper,we introduce the improved Pearson correlation coefficient and mutual information correlation analysis respectively to detect the dimensions' linear and non-linear correlations. For the linear case,all dimensions are classified into three groups according to their distributions.Then we correspondingly select the appropriate parameters for each group of dimensions to calculate their correlations.For the non-linear case,we cluster the data within each dimension.Then their probability distributions are calculated to analyze the dimensions' correlations and dependencies based on the mutual information correlation analysis.Finally,we use the relationships between dimensions as the criteria for interactive ordering of axes in parallel coordinate displays.

Keywords: `Multidimensional visualization`, correlation analysis, data distribution feature, dimension reordering

1. Introduction

The rapid development of information technology produces vast amounts of datasets with numerous dimensions and complex structures.These multidimensional datasets offer tremendous opportunities for studying behavioral patterns and predicting future developments. Valuable insight often comes from intricate inter-relationships that exist among data dimensions(or variables).However,for the data with many dimensions and complex structures,it is far from straightforwardly showing the relationships between dimensions in a meaningful and user-interpretable way.Traditionally,low-dimensional representations of high-dimensional spaces [1], obtained by methods such as Principal Component Analysis (PCA), Multi-Dimensional Scaling (MDS),Self-Organization Map(SOM),ect.are used to interpret their relationships from a macro perspective.Other methods mainly include the scatter plot matrix (SPM) and the parallel coordinate plot (PCP) can basically show some correlations between variables in the multidimensional data.

Correlation analysis is one of the most commonly used methods in multidimensional visualization.It looks for relationships between variables and can indicate whether the variables are related to each other and how strong of the dependency is.Pearson correlation coefficient[2] is a most commonly used method for correlation analysis which is proposed by Pearson in 1895.It is often used in the multidimensional visualization to directly characterize the correlations between two variables by the coefficient R.Another method called canonical correlation coefficient[3] is also often applied for the multidimensional visualization.Both of above methods use the correlation coefficient R to show the variables' linear correlations.However,they reflect inaccurate relationship when the datasets are not normal distribution.And they are easily influenced by the outliers.

Faced with the multidimensional data with a variety of distributions and structures,traditional linear correlation analysis methods are not efficient to analyze the data relationship.Therefore,we propose an improved method based on the Pearson correlation coefficient in this paper.We think that the calculation for Pearson correlation coefficient should use different parameters according to the datasets with different distributions.We first extract the statistical features of multidimensional data to judge each dimension's distribution.Then all dimensions are classified into three groups according to their distributions.Finally,we select the appropriate parameters to calculate the Pearson correlation coefficient for each group of dimensions.

Correlation coefficient is a good measure when the dimensions are nearly linear distributed.But it appears not suitable for the analysis of non-linear distributed dimensions in the multidimensional data.Furthermore, we propose a non-linear correlation analysis method based on mutual information correlation analysis [4] and clustering.We use the information entropy to measure the relationships between variables.It is assumed that the smaller the entropy,the stronger the relationship.On the contrary, the relationship is weaker.This method is not influenced by the distributions of datasets.It has a robustness for the noise points.Firstly,we divide the data within each dimension into some clusters.Then,the probability distribution is given by

Email addresses: `yizhang@tju.edu.cn` (Yi Zhang), `1012606185@qq.com` (Teng Liu*), `351540817@qq.com` (Kefei Li), `jwzhang@tju.edu.cn` (Jawan Zhang)