Visualization of High Dimensional Data

V

Gokmen Zararsiz
Hacettepe University Ankara, Turkey

Cenk Icoz
Anadolu University Eskisehir, Turkey

Erdener Ozcetin
Anadolu University Eskisehir, Turkey

INTRODUCTION

Data visualization is the computer aided mapping of data from numerical form to a Cartesian space to amplify cognition. It deals with using statistical tools called graphics, and these graphics include the combined use of points, lines, numbers, symbols, colors, etc. Over the years, data visualization has attracted greater attention from researchers than tabular representation due to the following advantages: (1) it is more efficient in attracting the interest of a reader, (2) more easy to understand and remember, (3) time saving (4) provides a more complete and better balanced understanding of a problem (Kromesch & Juhasz, 2005; Barthke, 2005; Meyer & Cook, 2000; Everitt & Hothorn, 2011).

Statistical algorithms are mostly based on the assumptions to be tested or certain rules to be obeyed. Generally, it is hard to follow the rules and to test whether these assumptions fit in or not. Therefore, the manipulation of data is sometimes required to fit the assumptions and for confirmation of the rules (Barthke, 2005; Keim, 2002). Chambers et al. (1983) stated the importance of data visualization as follows: "There is no statistical tool that is as powerful as a well-chosen graph" in (Chambers, Cleveland, Kleiner, & Tukey, 1983).

In statistical theory, high dimensional data (HDD) refers to data whose dimension is at least larger than the dimensions considered in traditional multivariate analysis. In many applications, it is possible to work with data in which the number of variables exceeds the number of samples. For example, gene expression data obtained by DNA microarray technology contain the gene expression levels of biological samples and the number of genes is usually in the hundreds to thousands where the actual number of biological samples is in the tens to hundreds (Cai & Shen, 2011; Tang & Zhang, 2002).

Over the last few years, significant developments in many fields have led researchers to work with HDD and it is now possible to store vast amounts of data in today's computers due to the advances in hardware technology. It is reported that every year, about an exabyte (one million terabytes) of data is generated, a large portion of which is available in digital form (Keim, 2002). However, extracting valuable information and relationships from HDD is a difficult and complex process and visualizing this data (with the help of statistical and data mining techniques) is a good way to simplify this complexity (Meyer & Cook, 2000; Keim, 2002).

In this chapter, we try to overview the most popular visualization techniques of HDD in detail. Furthermore, the classifications of these techniques are investigated and examples of graphics are given to the reader for further comprehension of the subject.

DOI: 10.4018/978-1-4666-5202-6.ch236

BACKGROUND

The visualization of data has always been a strong desire and an interesting application for data analysts (Fyfe & Garcia-Osorio, 2005). The era of data began after the pioneering works of John Wilder Tukey's in exploratory data analysis. Nowadays, not only statisticians and data analysts but also researchers in other fields such as doctors, chemists, and geologists aim to represent data visually in order to look for patterns and interactions. All these researchers try to find out the answers to questions which arise during their research by investigating only the data they have gathered (Donoho, 2000). As to be expected, data exist in all areas and will continue to increase.

During the 1970s, statistical graphics proposed for HDD were created to allow researchers to find patterns and interactions in progressively higher dimensions. Andrews' curves (Andrews, 1972) and Chernoff faces (Chernoff, 1973) are examples of these graphics. Dimension reduction techniques were also generalized to extract interesting information from HDD in lower dimensional graphics (Friendly, 2008). In 1974, PRIM-9, the first dynamic and interactive tool to view and manipulate HDD up to 9 dimensions, was developed (Fishkeller, Friedman, & Tukey, 1974).

In the early 1980s, more dynamic and specific visualization tools were introduced. Several examples of these tools include association plots (Cohen, 1980), mosaic plots (Hartigan & Kleiner, 1981) and sieve diagrams (Riedwyl & Schüpbach, 1986).

New methods for displaying data have been found and developed over the last two decades and they show great diversity today. These methods are proposed to determine outliers, recognize and identify patterns, diagnose models and generally to examine unexpected phenomena and different structures in data (Everitt & Hothorn, 2011).

MAIN FOCUS

One problem in the visualization of HDD is the limited visualization space. This is because the capability of the human brain is restricted and can only perceive up to three dimensional images. Also, this space is limited to two dimensions on paper or screens (Barthke, 2005). It has been stated that extracting valuable information by using dimension reduction techniques and visualizing it automatically causes problems when the number of variables increases. This is because the increasing number of dimensions will lead to the occurrence of a huge number of different data sets (Mihalisin, 2002).

There are several classifications and different classification schemes in HDD visualization in the literature. Kromesch & Juhasz (2005) proposed a classification for high dimensional data visualization techniques. They grouped these techniques into six broad categories. These techniques were: geometric, pixel-oriented, icon-based, hierarchical, graph based and hybrid techniques. Hybrid techniques were described as the combination of all the other techniques in the proposed classification. Another classification scheme was proposed by Bartke. Bartke (2005) classified HDD visualization techniques into three groups called graphical methods, icon-based methods and hierarchical methods. Similar classification schemes can be found in (Yang, 2005; Zhao, 2011). In all studies, visual examples are given for further comprehension. Table 1 displays the widespread classification of the widely used techniques. The use of methods differs depending on the main objective of the researcher (e.g. finding correlations, discovering clusters, detecting outliers, looking for patterns, etc.).