International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V72I8P115

Volume 72 Issue 8, 104-108, August 2024

Original Article

Visualizing Higher-Dimensional Data

Atmajitsinh Gohil

Independent Researcher, New York, United States of America.

Corresponding Author : atmajitsinh.gohil@gmail.com

Received: 18 June 2024 Revised: 22 July 2024 Accepted: 11 August 2024 Published: 30 August 2024

Abstract - *The advances in technology have led to generating not only large volumes of data but also at a higher frequency. The data generated is dynamic and available in different formats. The data generated in the present time is thus higher dimensional data. The primary objective of the paper is to review the data visualization techniques for summarizing and interpreting higher dimensional data. The paper studies the challenges of higher dimensional data and addresses the same through various visualization techniques.*

Keywords - *Data visualization, High dimensional data, Artificial Intelligence, Data analysis techniques, Data visualization tools.*

1. Introduction

Data is information that has been captured at a point in time, and as time evolves, the data points change, i.e. new information is made available with time. Recent advances in the field of computing power, processing and internet technology have led to the generation of larger data sets, generated at a higher frequency and in different formats. Large-scale data exists in various domains such as finance, healthcare, sports, environment, humanities and social sciences [1].

It is very crucial that the data collected is assessed and important information is extracted. The first step in processing large and unstructured data sets is through data visualization techniques. Data visualization provides techniques to summarize, interpret and communicate the information contained in large and complex data.

A dataset with one variable is called one-dimensional data and can be assessed through univariate analysis. A two-dimensional dataset comprises data arranged in a grid-like format with rows and columns and can be assessed through multivariate analysis. Higher dimensional data sets are complex and larger in size beyond typically considered in multivariate analysis. Higher dimensional data comprises data sets where the number of features is larger than the number of observations. Analyzing a complex and large data set is a trivial task; hence, data visualization techniques developed to assess univariate and multivariate data cannot be applied to evaluate higher-dimension data.

The present article provides an overview of various data visualization techniques that can be used for interpreting higher dimensional data. This paper also highlights the applications, extensions of visualization and limitations of the visualization techniques.

2. Challenges of Higher Dimensional Data

With the advances in technology, the data generated by various devices through social media or human interaction is predominantly high-dimensional data. High-dimensional data comes with challenges that are not evident in low-dimensional data, referred to as the curse of dimensionality. The datasets that are large, complex, dynamic and available in different formats give rise to challenges that are unique to high-dimensional data.

2.1. Processing Time

The processing power required for high dimensional data visualization is also high. Unlike static visualization, where the data is not dynamic, high-dimensional data is usually dynamic and frequently changing. Hence, the higher dimensional data requires additional processing power.

2.2. Rate of Image Change

The dynamic nature of high-dimensional data causes the visualization to change frequently. This problem becomes the most significant in monitoring tasks when a person who observes the data cannot react to the number of data changes or its intensity on display [3].

2.3. Visualization Loss

The primary objective of any visualization technique is to summarize data that is easy to interpret and extract useful information. Given the large dataset, the data analyst may fail to interpret the data visualization accurately unless the data is transformed. This problem comes from the fact that most of

the objects in the dataset are too relative to each other, and the screen watcher cannot divide them as separate objects.

2.4. Overview of the Tools

High dimensional data not only contains multiple variables but the dataset is large in size as well. Hence, it is important to understand the data visualization tools available along with understanding the pros and cons of each data visualization technique.

There are many data visualization tools available that can be employed. The data visualization tools can be broadly classified into Visualization applications, Programing Languages, Worksheets and Web-based programming languages.

2.4.1. Visualization Applications

Visualization applications such as Flourish, Qlik, and Tableau allow users to choose from a variety of visualization techniques. These applications also allow users to create interactive visualizations and dashboards that can be easily shared. These applications are easy to use but are not very flexible as the users are limited to the visualization templates available within the application.

2.4.2. Programming Languages

R and Python programming languages have many packages that can be used to create static, dynamic and interactive visualizations. The programming languages do not have pre-built visualization templates; hence, the users need to learn the programming language.

2.4.3. Worksheets

Google Sheets and Microsoft Excel can also be used to generate visualizations. However, the users are limited to the templates available, and as the size of the data increases, these tools are not as efficient as the other tools discussed in this section.

2.4.4. Web-Based Programming Languages

The visualization tool, such are D3.js uses Scalable Vector Graphics (SVG), HTML and CSS to create static and interactive visualizations which can be viewed in a web browser. The D3.js library provides the most flexibility; it requires users to understand HTML, CSS and JavaScript.

As discussed in his section, every visualization tool has its strengths and weaknesses. The users should evaluate each tool based on the user's requirements, data size, and data type to choose the best tool for their use.

3. High-Dimensional Data Visualization Methods

The present section describes data visualization methods that are used to visualize high-dimensional data. The methods are described along with the use case, pros and cons.

3.1. Chernoff Faces

Chernoff faces is a data visualization technique used to represent multivariate data. A human face features such as eyes, nose, mouth, and ears represent values of the variables by their shape, size, placement and orientation. The facial features, such as the length of the nose or face size, are used to represent up to 18 variables.

The advantage of using Chernoff's faces is that it enhances the user's ability to detect and comprehend important phenomena, serves as a mnemonic device for remembering major conclusions, communicating major conclusions to others, and providing the facility for doing relatively accurate calculations informally [4].

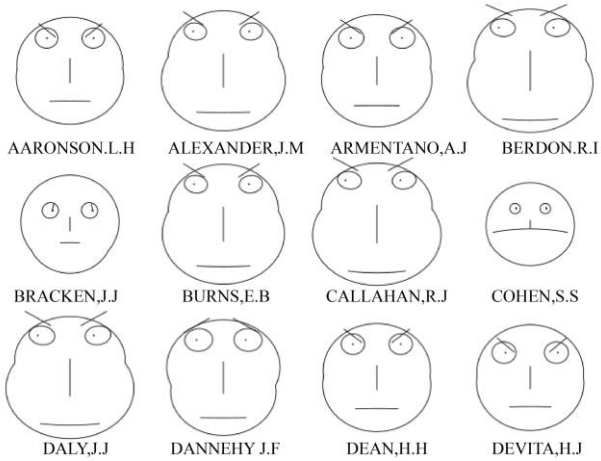

Fig. 1 Chernoff faces

The Chernoff faces in Figure 1 shows the judge's rating in the USA, where various features of the face represent variables such as judicial integrity, demeanor, preparation for trial, judicial integrity etc.

An extension of Chernoff faces the use of the Chernoff Emoji, where each feature of the Emoji represents the USA state ranking in a given metric. (Chernoff emoji)

3.2. Heatmaps

Heatmap provides a summary of multivariate data. A heatmap is a grid representation where each cell represents a data point, and the color of the cell represents the magnitude of the data. It is used to identify the trends in data, specifically areas of high values and low values. The advantage of using a heatmap is that it would allow for clustering and also allows for visualizing data with significantly high dimensions.

Heatmap visualization techniques have applications in various fields such as healthcare, climate change, finance, sports and demography. Heatmaps have been used in the healthcare sector, specifically in epidemiology studies, cancer research and pharmacovigilance. The heatmap visualization technique has been extended to create calendar heatmaps to identify patterns over time. Spatial heatmap has been used to

identify areas of high and low population density. In sports heatmap has been used to visualize basketball data. Website heatmaps are used to identify user engagement through scroll maps, click maps, and mouse-tracking heatmaps.

The heatmap in Figure 2 shows the USA crime rates in various states. The data shows crime in different offense categories such as Assault, Homicide, Human Trafficking, Kidnapping and Sexual offences committed against a person across various USA states in 2022. The visualization shows that all the states have a high assault rate (> 90%); however, Montana and Utah have low assault rates compared to other states. The visualization also identifies these 2 states with higher sexual offenses as well.

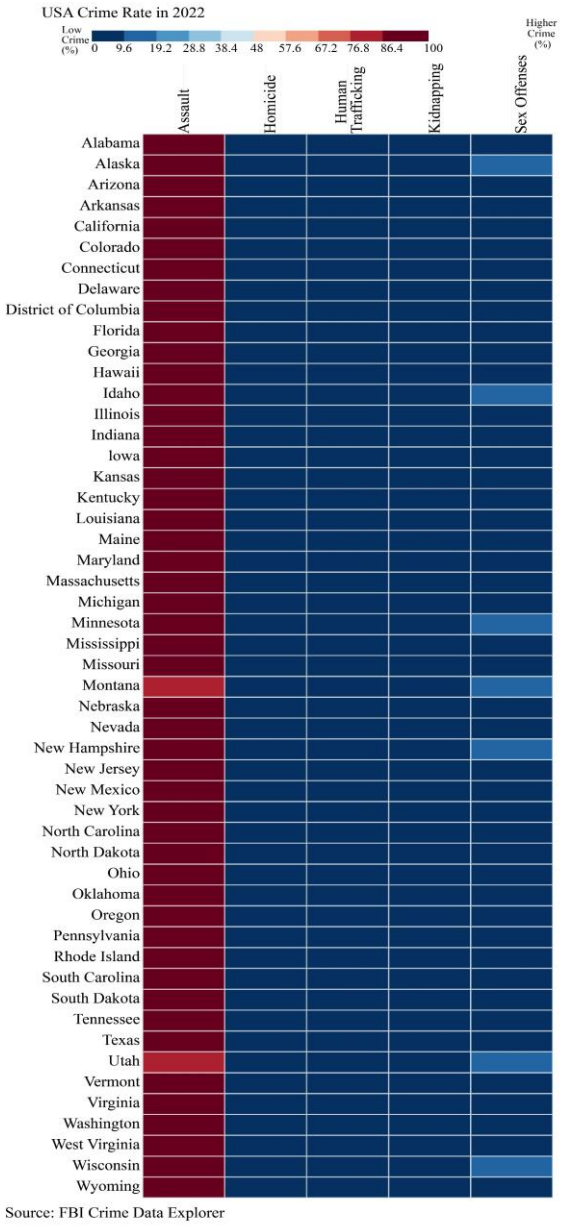

Fig. 2 Heatmap of USA Crime rate in 2022

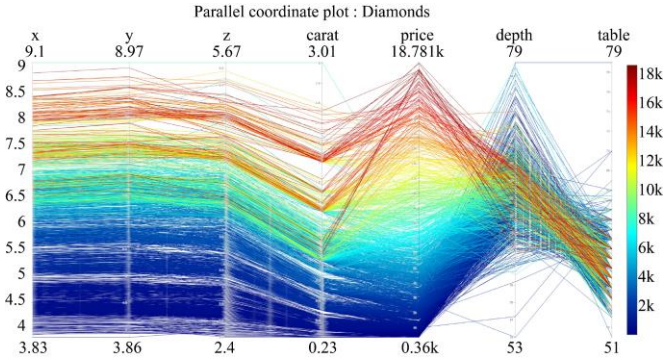

Fig. 3 Parallel plot showing the price and attributes of diamonds

3.3 Parallel Coordinate Plot

In the parallel coordinate plot, the variables are placed on the horizontal axis. Each vertical axis represents a variable that typically ranges from the minimum and maximum of that variable, so the highest value is plotted at the top and the lowest at the bottom [2].

Parallel coordinate plot has the advantage of plotting and analyzing all the data at once or individual data points and the ordering of variables on the horizontal axis does not impact the total diagram. The limitation of the parallel coordinate is that the number of variables that can be placed on the horizontal axis is limited and is not suitable for dynamic data [3].

The parallel plot in Figure 3 is created using the diamond data set. The dataset comprises 10 variables, and the parallel plot shows the relationship among 7 variables, i.e. X, Y, Z, Carat, Price, Depth and Table. The X, Y, and Z variables are the length, width and depth of the diamond in millimeters. The straight lines between X, Y and Z indicate a positive relationship. The variable carat (weight of the diamond) and price indicate a strong positive correlation (the majority of lines are straight), indicating that higher-carat diamonds are expensive. The lines cross each other between the price of a diamond and the depth (depth percentage) of a diamond, indicating a weak correlation.

The diamond data comprises 53,940 observations; hence, it is difficult to interpret the relationship using Figure 3, where the entire data is used to create the parallel plot. The visualization can be improved by highlighting diamonds above a certain threshold and using grey color for all other diamonds. Additionally, the visualization can be made interactive to allow users to order the variables or choose colors that indicate a positive and negative correlation.

3.4. Tree Maps

Treemaps are rectangles placed together to form a grid wherein the size of each rectangle is directly proportional to the data it represents. A higher value would represent a bigger

rectangle. A portfolio manager often uses tree maps to show the returns from different sectors of the portfolio.

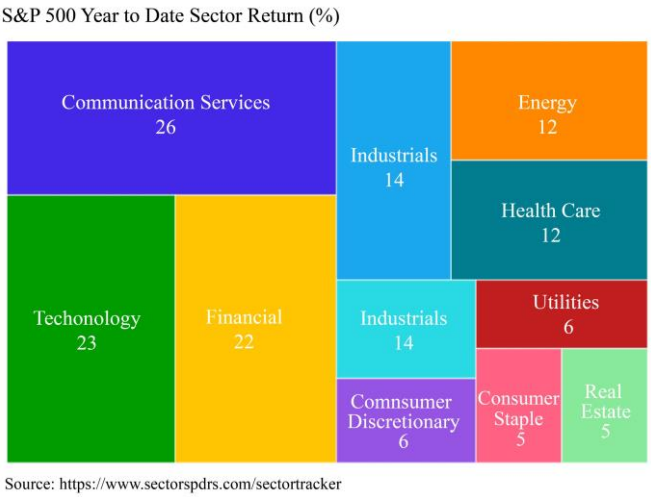

Fig. 3 Tree Map of S&P 500 Sector returns

Treemaps can be used to show an increase or decrease in data or analyze relationships. The diverging color scheme can be used to fill the rectangles of treemaps to compare data between two periods. A sequential color scheme can be used to analyze the data [5].

The primary advantage of a tree map is that data relationships can be shown based on hierarchical grouping, and extreme outliers can be easily detected using special colors. The tree maps suffer from the weakness that it is not suitable for time series or observing historical trends and the variable used to create the size of the rectangle cannot be a negative value. Circle packing and Sunburst are alternative visualization techniques that can be used instead of a tree map but they suffer from the same limitations as a tree map.

Figure 3 shows the tree map of different sectors of the S&P 500. The Communication Services, Technology, and Financial sectors have the highest return. The Real Estate, Consumers Staples and Consumer Discretionary have the lowest returns.

The portfolio or financial indices also include subsectors under each sector. The subsectors can be shown by dividing the sector into smaller rectangles.

3.5. Star Plot

In a star plot, each data item is portrayed by a one-star symbol, which is positioned in a grid. The rays of the star (axes), which are equally spaced around a circle, represent different variables of a data item.

The star plots representing each data item are placed in a grid; hence, it allows the users to compare the shape of one data item with another.

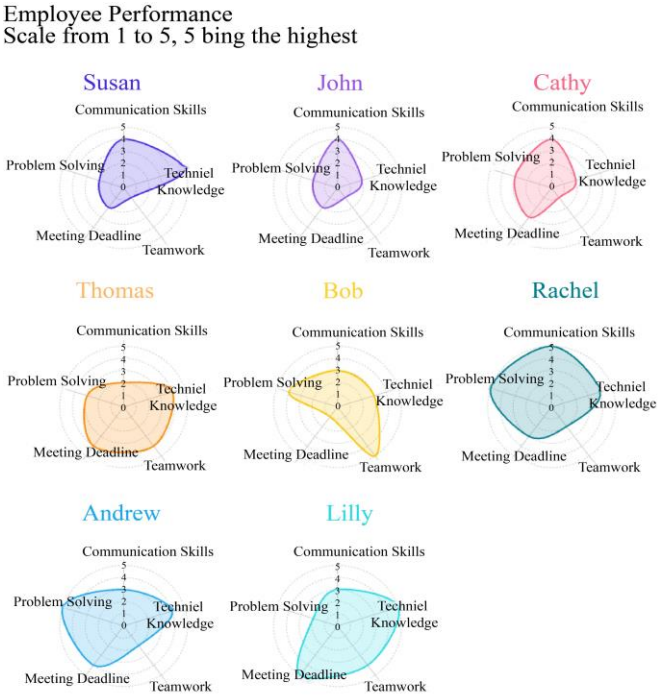

Fig. 4 Radar plot

Figure 4 shows the radar plot of 7 hypothetical employees' performance. These employees are compared using categories such as communication skills, problem-solving, meeting deadlines, teamwork, and technical knowledge. The employees are ranked from 1 to 5, with 5 being the highest. The Radar plot is a great tool to perform comparative analysis, e.g. John, Susan and Bob are low performers in 3 of the 5 categories compared to other employees at the firm.

3.6. Scatterplot Matrix

A scatterplot is used primarily to represent relationships between two variables. When the number of variables is large, a scatterplot matrix can be used. Scatterplot matrices can be used as an alternative to parallel coordinate plots and they are easier to interpret as well.

A scatterplot matrix visualization comprises creating a grid of scatterplots. This technique allows the users to visualize all the variables and their relationships. The scatterplot matrix can also be combined with other plots, such as a histogram, to show the distribution of the variable.

Figure 5 shows the scatterplot matrix of the diamond data set. The plot compares the relationship between variables such as price, carat, depth, and table. The lower part of the plot shows the relationship between variables, and the diagonal shows the density plot. The upper part of the plot can be used to display another set of plots, such as a box plot or correlation value.

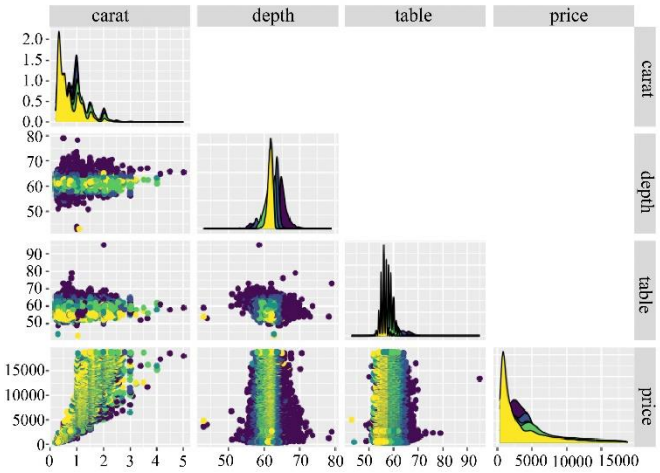

Fig. 5 Scatterplot matrix

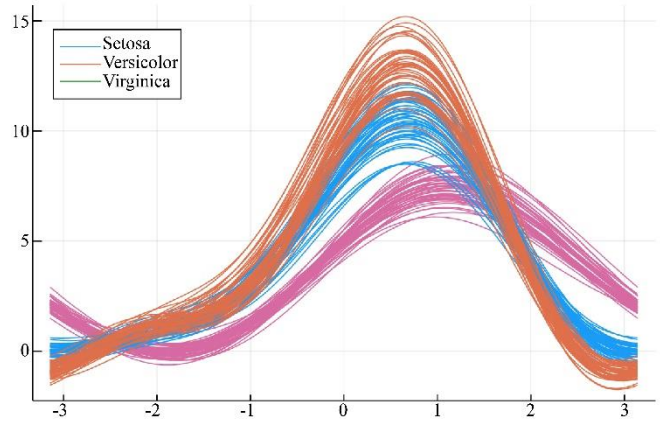

Fig. 6 Andrews plot of iris data

3.7. Andrews Plot

The Andrews plot is les renowned but highly effective technique to visualize data in higher dimensions. The Andrews plot reduces higher dimensional data to lower dimensional data by retaining the relative distance between other samples and keeping the variance similar.

If the data is k-dimensional, each point $X` = (X_1 \dots X_k)$ defines a function:

$$f_x(t) = \frac{X_1}{\sqrt{2}} + X_2 \, Sin\, t + X_3 \, Cos\, t + X_4 \, Sin\, 2t + X_5 \, Cos\, 2t + \cdots$$

This function is then plotted over the range $-\boldsymbol{\pi} < \boldsymbol{t} < \boldsymbol{\pi}$.

The Iris data comprises of Sepal length and Width and Petal Length and Width of 50 irises from each of the 3 species: Setosa, Virginica and Versicolor. From the above plot, it is clear that the Setosa is separate from the other 2 species. The curves are tightly banded near t0 = -2.5. This implies that, in a direction orthogonal to the vector the data cloud in R^4 looks flat, so a dimension reduction from 4 to 3 is likely. Further, the Versicolor species has potential outlying curves in the range of 0 to 1[6].

$$(\mathbf{1}/\sqrt{\mathbf{2}}, \boldsymbol{Sin\ t0}, \boldsymbol{Cos\ t0}, \boldsymbol{sin\ 2t0})$$

4. Conclusion

The present paper aims to review the challenges of higher dimensional data and review data visualization techniques that can be employed to assess higher dimensional data.

The tools used for interpreting and summarizing high-dimensional data play a pivotal role as the volume of data increases with the advancement of AI and ML. The higher dimensional data brings along challenges that were not evident with low dimensional data. Hence the visualization methods used should aim to address these challenges. There is no one visualization tool or technique that can be applied to higher dimensional data. Hence, the users should understand the advantages and limitations of each visualization prior to processing the data.

Future work in the field of data visualization can be in the areas of developing visualization techniques that can address dynamic data, utilizes low computation power and can be generated quickly. The high dimensional data comprise many variables; hence, future research should aim at developing a framework that combines both data transformations and visualizations.

References

[1] D. Srivastava, "An Introduction to Data Visualization Tools and Techniques in Various Domains," *International Journal of Computer Trends Technology*, vol. 71, no. 4, pp.125-30, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Nathan Yau, *Data Points: Visualization that Means Something*, John Wiley & Sons, 2013. [Google Scholar] [Publisher Link]

[3] Evgeniy Yur'evich Gorodov, and Vasiliy Vasil'evich Gubarev, "Analytical Review of Data Visualization Methods in Application to Big Data," *Journal of Electrical and Computer Engineering*, vol. 2013, no. 1, pp. 1-7, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[4] Herman Chernoff, "The Use of Faces to Represent Points in K-Dimensional Space Graphically," *Journal of the American Statistical Association*, vol. 68, no. 342 pp. 361-368, 1973. [CrossRef] [Google Scholar] [Publisher Link]

[5] Martijn Tennekes, and Edwin de Jonge, "Top-Down Data Analysis with Treemaps," *In Proceedings of the International Conference on Imaging Theory and Applications*, Vilamoura, Algarve, Portugal, vol. 2, pp. 236-241, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[6] Paul Embrechts, and Agnes M. Herzberg, *Variations of Andrews' Plots*, International Statistical Review / Revue Internationale de Statistique, vol. 59, no. 2, pp. 175-194, 1991. [CrossRef] [Google Scholar] [Publisher Link]