Parallel Coordinate Plots

Parallel coordinate plots are a data visualization technique which can be used to analyze multivariate numerical data. Multivariate data refers to data which involves two or more variable quantities.

Using charts and graphs to highlight data and relationships between data sets is often advantageous as they serve to enhance our understanding of behavior and trends, in relation to different variables. But in many cases, these co-relations and relationships are highly multidimensional in nature, and we tend to typically use low-dimensional cross sections of the data to convey the relationships. To give a more insightful look into the data and how the variables effect it, data analytics is utilized together with visualization strategies to better represent the multidimensional nature of the data. These are parallel coordinate plots.

With the amount of data generated today, effective tools are required to analyze and present them in an effective manner, without losing sight of the overview the data represents. Informative data visualization allows us to explore and understand large data sets in an intuitive manner. This is where the technique of parallel coordinate plots come in- to analyze multivariate data and draw intuitive conclusions. Some examples of actual applications of parallel coordinate plots include monitoring and predicting weather conditions, analyzing electroencephalography (EEGs) for finding relevant differences in patients and healthy people, gene experiments to assess differences, and other situation which have several sets of variable data.

When using parallel coordinates, individual data elements are plotted across several performance measures. Each of these performance measures corresponds to a vertical axis and each of the data elements is then displayed as a series of connected points along the axes. Parallel coordinate plots are generally used when of samples or observations with multiple variables ne ed to be compared. We can take the simple example of a laboratory which needs to measure the amount of various carbohydrates- glucose, fructose, maltose, saccharose- in different fruits and vegetables. The variables, in this case the different carbohydrates, will be shown as vertical axes. A parallel coordinate plot in this case allows the comparison of how the carbohydrate distribution varies from fruit to fruit or vegetable, and what similarities can be observed.

Invented by Alfred Inselberg in the 1970's as a way of visualizing high-dimensional data, parallel coordinate plots are typically found only in academic and scientific communities and not much in business or consumer data visualizations. This is because parallel coordinate charts often become very dense and difficult to understand- for the layman at least. But they are a godsend for people who work with data. This is so because the main strength of the parallel coordinate is in their ability to bring forward and reveal meaningful patterns and comparisons when used interactively for analytics. When interactive highlighting, roll-over details and filtering are applied, parallel coordinate charts can reveal some interesting information in your data.

If Parallel Coordinate Charts Are So Complicated, Why Use Them?

With most standard charts, there is a limitation on how many measures you can effectively plot or show. Standard progression of charts by measures are typically as follows:

- two measures: Scatterplot charts
- three measures: Bubble charts

- four measures: Bubble charts with colors
- five measures: Bubble charts with colors and animations

And with five measures on a single chart, you have effectively made an almost indecipherable graphic. That's where parallel coordinates find their niche- in showing many measures, limited only by horizontal space. They are therefore perfect for comparing several numerical variables simultaneously when those variables have different scales or magnitudes and different units of measurement. The key idea is to find similarities, patterns, clusters, positive relationships, negative relationships, or even no relationship, in multidimensional datasets.

How Are Parallel Coordinate Plots Used?

Each numerical value is a parallel coordinate chart Id given its own axis. All axes are placed parallel, vertically, and equally spaced. Every data element of the data set in use is represented through line segments which are connected. These are derived from a connected set of points, one on each axis of the chart. When this is done, we finally get a set of lines, each of which denotes a multi-axis representation of each data record. The general hypotheses and understandings are that several parallel lines indicate a positive relationship and lots of lines crossing indicate an association which leans towards the negative.

normalizing data for parallel coordinate plots

An important point to note is that the data points for a parallel coordinate chart have to be normalized. Data normalization, in simple words, is the process of reorganizing data in a proper manner. It is a data design technique that serves to reduce redundancy and eliminates undesirable features of your data such as update and deletion anomalies or insertion. Data normalization rules will divide larger tables into smaller tables and will link them using relationships.

In the case of parallel coordinate plots, normalization must take place before the data is plotted, as it is important to note that the raw data may have different orders of magnitude as well as different units of measurement. Normalization or scaling will transform the raw data into a new scale which will allow the comparison of different values which were initially of different magnitudes.

There are several normalization techniques available. The usual practice is to scale each axis between its minimum and maximum values. This technique, although not very robust, is the most widely used. The lowest value is set to 0 and highest to 100 per cent (or 1), and all other values in between are thereafter transformed accordingly. Other scaling techniques make use of the mean and standard deviations, the median or any other statistic to transform the original values to a common scale.

Apart from scaling, three other techniques are generally used in a parallel coordinate plot to discover patterns and relationships across variables. These are coloring, reordering and brushing. We take a look into these three briefly:

Coloring

This implies the highlighting of an individual line or a set of line by using a particular color. This visually separates them from the others. An example where coloring can be used is when you have to compare one particular product against the rest of the products on the list.

Reordering

Reordering involves changing the order of the vertical axes. Reordering is done because relationships between adjacent variables are more easily visualized than between non-adjacent variables. There are high chances of missing out on identifying patterns and relationships simply because the variables are not located adjacent to each other on the chart. Also, the reduction of clutter of the parallel coordinate plot can be reduced by changing the order of some of the axes and minimizing the crosses between them.

The great thing about modern visualization tools is that they allow the axes to be dragged along the plot to facilitate the reordering. A good analyst will always experiment with reordering until the readability of the plot is enhanced to the maximum and as much information as possible can be ascertained from the display.

Brushing

Brushing is a technique in which individual data points are selected by the application of a brush in order to highlight subsets of the data in question. Some lines will be emphasized will others are faded. Think of the brush like a filter that reduces the number of lines and minimizes clutter, thereby revealing patters in the dataset in the process. Brushing is a mandatory activity for parallel coordinate plots and charts as it negates over-plotting and occlusion during the analysis of larger data sets.

Why Use Parallel Coordinate Plots?

When compared to other data analysis techniques, parallel coordinate plots come on top, especially when it comes to exploring and visualizing multivariate data – or a dataset with many variables. But of course, there are more than one way of exploring multivariate data. It is therefore important to evaluate or investigate which technique would work best in different conditions.

As highlighted earlier, the foremost advantage of the parallel coordinate plot is its ability to represent high dimensional data as a two-dimensional visualization. This means that as the data is represented in the form of a line, it becomes easier to pick out trends and identify patterns. Additionally, the axes representing the variables can be moved around to explore and compare the trend between different axes. This allows an in-depth insight which may not be possible in other data visualization techniques.

What To Watch Out for When Using Parallel Coordinate Plots?

We have just listed down how parallel coordinate plots are excellent for visualizing data across multiple measures. Why then, is it not the most popular data visualization technique? Here are some of the issues with parallel coordinate plots:

- Large data sets will create a lot of clutter when visualized. Which means that sometimes, meaningful patterns and trends can be obscured in the clutter of several lines overlapping or crisscrossing each other.

- The order of how the axes is laid out will impact how the reader will understand the data. Relationships between adjacent axes or measures are easier to perceive which relationships between non-adjacent measures may not be easily discernable.
- As the axes keep getting closer to each other, it becomes more challenging and difficult to perceive the structure or clusters.
- As each axis is dependent on different data variables, each axis can have a different scale. This can become difficult to display as well as difficult for the reader to absorb.
- Lines can sometime be mistaken for trends/patterns or even changes in value; despite being used only to show the connected relationships between points.
- The data needs to be numerical as parallel coordinate plots simply do not work well for categorical data or those which have few values per axis.
- Furthermore, there are limitation to what can be done with parallel coordinate plots. When the number of data items increases, overplotting results. This overplotting often makes it next to impossible to see anything. The number of dimensions on the screen also must below a dozen concurrently to really make sense. Anything above this number and the visualization gets very challenging to read.
- Parallel coordinate plots present challenges in that they are difficult to understand for non-technical audiences.
- One has to show a limited number of numerical values at a time as, as mentioned above, more than twelve axes and their corresponding lines could confuse the audience. The technique of brushing may not be able to limit this either.
- Scaling means that the original value of every variable is lost.

Challenges made note of, there are quite a few ways in which most of these can be met. These solutions include the reordering of axes, clustering of similar axes and others. But the fact remains that parallel coordinate plots works best for datasets which have a moderate number of dimensions and the number of records in it are no more than a few thousand.

In Conclusion

Visualization of information to understand the semantics and intricacies of data is a necessary activity today. Considering the reliance on data for almost all activities- from business, education, shopping to eating- the need to visualize, interact and then analyze data to extract meaningful patterns and relationships, will only increase in the future. Data today is growing exponentially, and data visualization techniques are taking on an increasingly important role in several areas such as e-commerce, medical research, weather research and prediction and many more. Parallel coordinate plots can play an important role in these areas.

Parallel coordinate plots may look frightening to a non-technical person or even to those who have limited technical knowledge but are in reality quite approachable. Parallel coordinates are versatile and thus a very useful tool for finding structure and patterns in moderately sized data sets. With some experience and practice it is possible to quickly recognize patterns in data sets or even estimate the strength of correlations between the axes. The flexibility of parallel coordinate plots in allowing the user

to easily manipulate and rearrange the axes to highlight important lines via brushing and coloring or even filtering data in an interactive manner, allows for an extremely versatile data visualization tool.