High Dimensional Data Visualization

Sándor Kromesch, Sándor Juhász

Department of Automation and Applied Informatics, Budapest University of Technology and Economics, Budapest, Hungary
Tel.: +36-1-463-3969; Fax: +36-1-463-3478
E-mail: sandor.kromesch@aut.bme.hu, sanyo@aut.bme.hu

Abstract: Visualizations that can handle flat files, or simple table data are most often used in data mining. Over the last few years many techniques have been developed for visualising different types of information. A brief background to data visualization and key references are provided in this paper. Our goal is to review the various high-dimensional visualization techniques and classify and summarize these in a comparative table in which we emphasize the main properties of these methods.

Keywords: visualization techniques overview, high dimensional data visualization

1 Introduction

Visualization is visual representation of data. data are mapped to numerical form and translated into graphical representation. The simple line graph or scatter plot has been used for visualization for hundreds of years. Perhaps they are the most widespred method of understanding the interaction of two variables. Understanding the expression of one value as a function of the second is easier if the function is plotted on a graph. The relationships between three variables can be partially understood by a three-dimensional view. The ability to understand the interactions or correlations between the more than three variables becomes severely limited using standard visualization tools. The area of information visualization and its associated areas of data mining are relatively new and need different visualization tools as the scientific visualization for example. The focus of this article is presenting the techniques of information visualization.

2 Classification of Data Visualization Techinques

Data visualization is graphical presentation of a data set with the aim of providing viewers a quality understanding of the information contents in natural and direct way. Direct display of data with more than three dimensions is difficult, which means that users should understand that the display of data with more than three dimensions has to be transformed in some way before they can be rendered. Users have to be aware of this transformation, and be able to reverse the transformed display and restore the original picture in their mind. The most popular visualization techniques are classified in geometric, icon-based, pixel-oriented, hierarchical, graph-based, or hybrid class. The Table 1 shows the widespread calssification of data visualization techniques.

Classes	Techniques
Geometric	Scatterplots [1, 2], Lanscapes [3], Projection Pursuit [4],Prosection Views [5,6], Hyperslice [7], Parallel Coordinates [8, 9]
Icon-Based	Chernoff Faces [10, 11], Stick Figures [12, 13], Shape.Coding [14], Color Icons [15, 16]
Pixel-oriented	Recursive Pattern Technique [17], Circle Segment Technique[18], Spiral- and Axes Techniques [19]
Hierarchical	Dimensional Stacking [20], World-within-World [21], Treemap [22, 23], Cone Trees [24], InfoCube [25]
Graph-Based	Basic Graph (Straight-Line, Polyline, Curved-Line) [26]
Hybrid	Arrbitrary combination from above

Table 1
Classification of Data Visualization techniques

3 Basic Techniques

We use some basic techniques to transform data sets before visualization, because the types of the data in datasets are very various. On the other hand the visualization techniques require some structure of data. For this reason we have to transform our data set before the vizualization.

The one of the widespread procedures the data set is got over on is the normalization. First normalization technique is the **local normalization** (LN) in which each dimensional range is scaled to be between 0 and 1, to give each dimension equal weight. In **global normalization** (GN) all data values are scaled to be between the maximum and minimum of all data values. The dimension that

has the highest data values may have greater weight or impact in some visualizations. We prefer to scale the data to a range between 0 and 1 or -1 and 1, because it is more advantageous for some visualizations or data mining algorithms. In most cases local normalization is used for visualization. When many columns of a dataset are the same type, it is favourable for those columns to have the same normalization. For the full flexibility, it to be ensured the capability to transform each column with a different "weighted" normalization.

The technique where data points in a visualization are randomly moved a small distance to improve the display or to reveal obscured (overlapped) data points. This techique is called **jittering or agitating** [27]. Splatting is a technique for allowing three-dimensional scatter plot points to have some transparency and get more points on the screen.

Dimension brushing is a technique for highlighting a particular n-dimensional subspace in a Table Visualization. Each dimension has an active range, and the points within this active range can be coloured or highlighted with a particular value. The visualization area of this brush is colored differently.

Colour is an important and frequently used feature in data visualization. In multidimensional data visualization, for example in 4D, we can use the first 3 dimensions to construct a set of 3D objects, then we use different colour for the points to indicate the difference in the fourth dimension data.

Similarly to the colour, **size** as an attribute of data point is very important in data visualization. In the sane way, the data point can be either in a dimensional plane, or in a 3 or higher dimensional space. There are several drawbacks of size which include the following: the first is that it maybe difficult to implement in conjunction with any projection that introduces a decrease in size with apparent distance. Secondly, the size of each data point will greatly reduce the total number of data points in one view.

Glyphs (also referred to as icons) are graphical entities which convey one or more data values via attributes such as shape, size, color, and position. The use of glyph attributes for data visualization is based on human perceptual abilities.

The typical datasets for the visualization examples are either the automobile or the Iris flower dataset. Nearly every data mining package comes with at least one of these two datasets. The datasets are available UC Irvine Machine Learning Repository [28].

4 Methods

In this paragraph, we discuss the nine widespread visualization techniques. These techniques are estimated in accordance with two viewpoints. The first viewpoint is

the usability of the diagramms and the second is the computation demand on the use of techniques. The criteria of usability are defined in the following list:

- See correlation
- See outliers
- See clusters
- See rules and patterns
- See important features

The second viewpoint is computation demand on the visualization method.

Technique	C. demand	Usability
Line graph	small	small
Scatterplot matrix	medium	great
Bar cahrt	small	small
Permutation matrix	small	medium
Survey plot	small	medium
Parallel coordinates	small	medium
Circular parallel coordinates	small	medium
Andrew's curves	small	small
Radviz	great	great

Table 2
Comparison of Visualization techniques

4.1 Line Graphs/Multiple Line Graphs

Line graphs are used for displaying single valued or piecewise continuous functions of one variable. They are normally used for two-dimensional data (x, y) where the x value is not repeated (x is time variable, for example). A background grid can be used to help determine the absolute values of the points. Multiple line graphs can be used or overlaid to show more than two dimensions. (x, y1,y2,y3……) The fact that the first dimension or the independent variable is unique gives this dimension special significance. This dimension typically represents the ordering of the table (a number from 1 to M). Often this initial ordering of the table is correlated to one of the dimensions of the data, such as time. There are problems with using overlaid multiple line graphs. Different types of continuous lines (colored, dashed, etc) have to be used to distinguish each dimension. Each dimension may have a different scale that should be shown. For

more than three dimensions, multiple line graphs can become confusing, depending on the scale and whether or not an offset is used to separate the dimensions. If different colored lines are used to identify each dimension, brushing with color can not be used to distinguish classification points. For noting interdimensional correlation, a Survey Plot might be more suitable. If one is analyzing one or two continuous functions in detail, a line graph is very appropriate. Figure 1 is a multi-line graph of the Car dataset, where the X-axis orders the data records by type (American, Japanese, and European). Within each type, the records are sorted by year. The Y-axis represents the various dimensions (locally normalized) and the line plots are colored by the type of car. The offsets between line plots are the same because the range of each line plot is the same after the normalization.

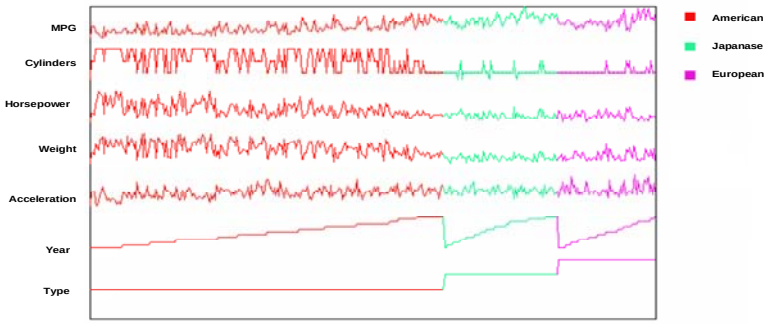

Figure 1
Multi-Line Graph of the Car Dataset

4.2 Scatter Plot Matrix

A Grid of two-dimensional scatterplots is the standard way of extending the scatter plot to higher dimensions. If one has 10 dimensional data, a 10 X 10 array of scatter plots is used to provide a visualization of each dimension versus every other dimension. This is useful for looking at all possible two-way interactions or correlations between dimensions. The standard display quickly becomes inadequate for high dimensions and user interactions of zooming and panning are needed to interpret the scatter plots effectively. Figure 2 shows a scatter plot matrix for the Car dataset. Data for American cars is red, Japanese green, and European blue. Positive correlations can be seen between Horsepower and Weight. Negative correlations can be seen between MPG and Horsepower and Weight.

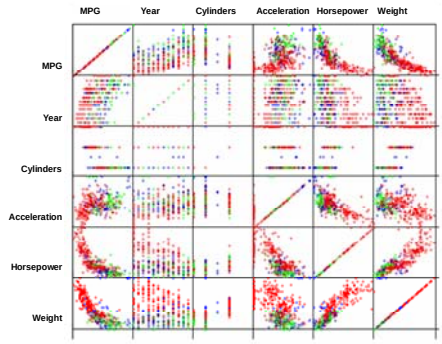

Figure 2
A Scatter Plot Matrix of the Car Dataset

4.3 Bar Charts, Histograms

Bar Charts are normally used for presentation purposes. They are related to histograms and Survey plots, which are used frequently in data mining. Bar charts are line graphs with the area under the line filled in. Data points usually repeated to widen the bar. Histograms are bar charts (or graphs) where the value for the bar represents a sum of data points. Histograms visualize discrete probability density functions. Multiple bar graphs and histograms can be used effectively (see Survey Plots) in data mining. One can use an array of histograms to approximate the density functions of all the dimensions of the data. Figure 3 is a histogram matrix of the Car dataset dimensions for each class.

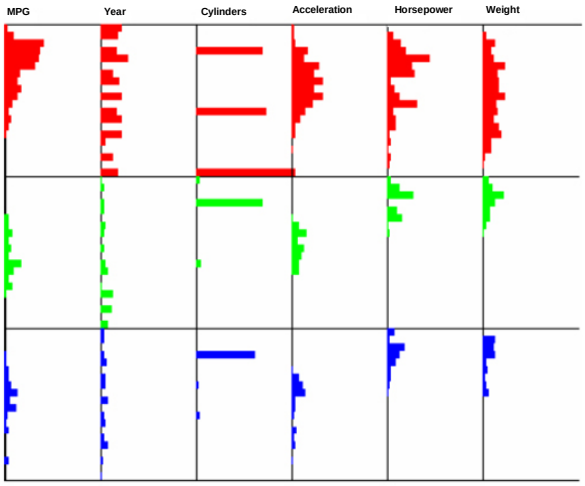

Figure 3
A histogram matrix of the Car dataset

4.4 Permutation Matrix

In the 1960's, Jacques Bertin [29] introduced the Permutation Matrix. This interactive plot is very similar to the Survey Plot, whereby the heights of bars correspond to data values. By permuting or sorting the rows or columns depending how the data is oriented in some manner, patterns in the data can be seen quite easily. In one of the modes of the Permutation Matrix, all data values below the average value are colored black and all data values above average are colored white. A green dashed bar corresponding to the average data value for each dimension is also displayed through the data.

4.5 Survey Plots

A simple technique of extending a point in a line graph (like a bar graph) down to an axis has been used in many systems. A simple variation of this extends a line from a center point, where the length of the line corresponds to the dimensional value. This particular visualization of n-dimensional data allows one to see correlations between any two variables especially when the data is sorted according to a particular dimension. When color is used for different classifications, one can sometimes see (using a sort) which dimensions are best at classifying data. The survey plot in Figure 4 shows American (red), Japanese (green) and European cars. The data is sorted by cylinders and then by miles per gallon.

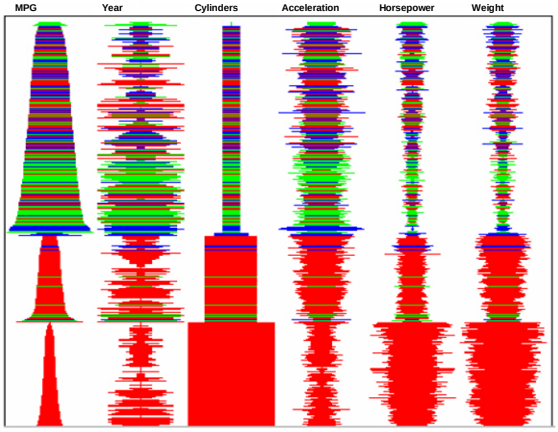

Figure 4
A Survey plot of the Car Dataset sorted by Cylinders and MPG

4.6 Parallel Coordinates

Parallel Coordinates represents multidimensional data using lines. A vertical line represents each dimension or attribute. The maximum and minimum values of that dimension are usually scaled to the upper and lower points on these vertical lines. N-1 lines connected to each vertical line at the appropriate dimensional value represent an N-dimensional point. In Figure 5 the Iris flower data is displayed using Parallel Coordinates where the three types of Flowers are represented with red, green and purple lines.

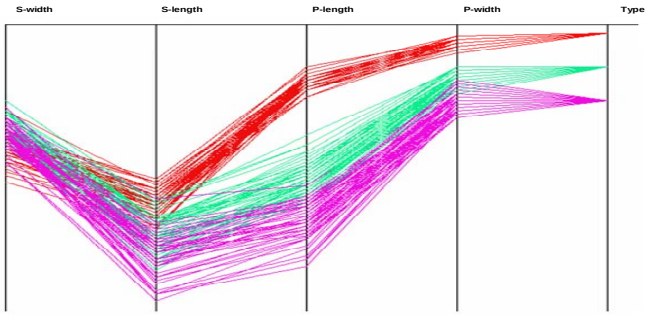

Figure 5
Parallel Coordinates of the Iris dataset (GN)

4.7 Circular Parallel Coordinates

A simple variation of Parallel Coordinates is a circular version, in which the axes radiate out from the center of a circle and extend to the perimeter. The line segments are longer on the outer part of the circle where higher data values are typically mapped, while inner dimensional values toward the center of the circle are more cluttered.

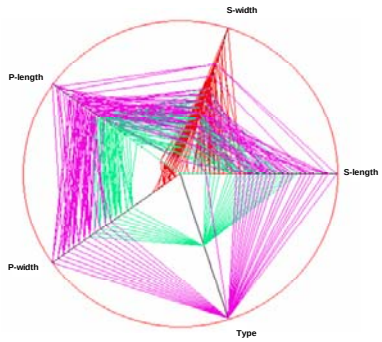

Figure 6
Circular Parallel Coordinates of the Iris Flower dataset (LN)

This visualization is actually star glyphs (plots) of the data superimposed on each other. Because of the asymmetry of lower (inner) data values from higher ones, certain patterns may be easier to detect with this visualization. Figure 6 is a Circular Parallel Coordinates display of the Iris flower dataset.

4.8 Andrew's Curves

Andrews' curves plot each N-dimensional point as a curved line using the function

$$f(t) = \frac{x1}{\sqrt{2}} + x2*\sin(t) + x3*\cos(t) + x4*\sin(2t) + x5*\cos(2t) + \ldots \quad , \qquad (1)$$

where the n-dimensional point is X = (x1,x2,….., xn) The function is usually plotted in the interval - $P_i < t < P_i$. This is similar to a Fourier transform of a data point. One advantage of this visualization is that it can represent many dimensions. A disadvantage is the computational time to display each n-dimensional point for large datasets. In Figure 7 the Iris flower dataset is plotted using Andrews' Curves.

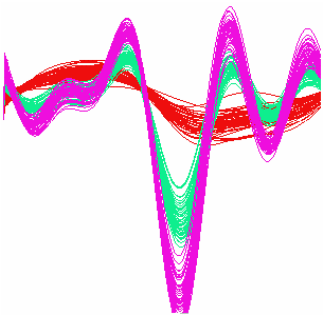

Figure 7
Andrews' Curves of Iris dataset (3 types of flowers)

4.9 Radviz

Spring constants can be used to represent relational values between points [30],[31] developed a radial visualization (Radviz), similar in spirit to parallel coordinates (lossless visualization), in which n-dimensional data points are laid out as points equally spaced around the perimeter of a circle. The ends of each of n springs are attached to these n perimeter points. The other ends of the springs are attached to a data point. The spring constant K_i equals the values of the i-th coordinate of the fixed point. Each data point is then displayed where the sum of the spring forces equals 0. All the data point values are usually normalized to have values between 0 and 1. For example if all n coordinates have the same value, the data point will lie exactly in the center of the circle. If the point is a unit vector,

then that point will lie exactly at the fixed point on the edge of the circle (where the spring for that dimension is fixed). Many points can map to the same position. This represents a non-linear transformation of the data, which preserves certain symmetries and which produces an intuitive display. Some features of this visualization are:

- Points with approximately equal coordinate values will lie close to the center
- Points with similar values whose dimensions are opposite each other on the circle will lie near the center
- Points which have one or two coordinate values greater than the others lie closer to those dimensions
- An n-dimensional line will map to a line
- A sphere will map to an ellipse
- An n-dimensional plane maps to a bounded polygon

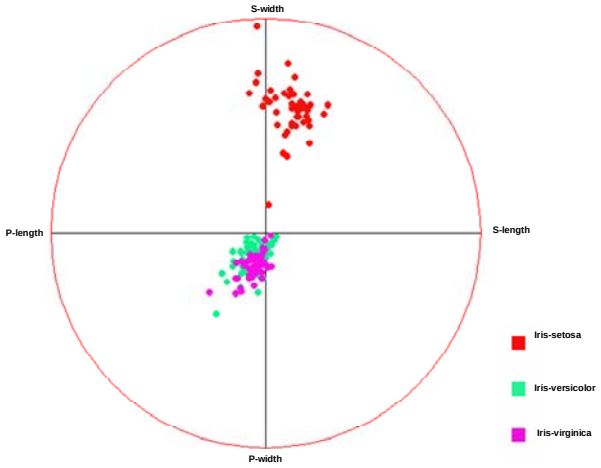

Figure 8
Circular Parallel Coordinates of the Iris Flower dataset (LN)

Conclusions

In this paper we introduced the most frequented techniques used in data visualization. We reviewed the basic procedures of data sets used before data visualization, and classified the the used technics. We stressed the advatntages and disatvantages of these methods and nine technics were detailed in this paper.

Acknowledgement

This work was supported by the Mobile Innovation Center, Hungary.

References

[1] Andrews D. F.: Plots of High-Dimensional Data, Biometrics, pages 125-136, March 1972

[2] Cleveland W. S.: *'Visualizing Data'*, AT&T Bell Laboratories, Murray Hill, NJ, Hobart Press, Summit NJ, 1993

[3] Wright W.: *'Information Animation Applications in the Capital Markets'*, Proc. Int. Symp. on Information Visualization, Atlanta, GA, 1995, pp. 19-25

[4] Huber P. J.: *'Projection Pursuit'*, The Annals of Statistics, Vol. 13, No. 2, 1985, pp. 435-474

[5] Furnas G. W., Bederson B. B.: *'Space-Scale Diagrams: Unterstanding Multiscale Interfaces'*, Proc. Human Factors in Computing Systems CHI '95 Conf., Denver, CO, 1995

[6] Spence R., Tweedie L., Dawkes H., Su H.: *'Visualization for Functional Design'*, Proc. Int. Symp. on Information Visualization (InfoVis '95), Atlanta, GA, 1995, pp. 4-10

[7] van Wijk J. J., van Liere R. D.: *'Hyperslice'*, Proc. Visualization '93, San Jose, CA, 1993, pp. 119-125

[8] Inselberg A.: *'The Plane with Parallel Coordinates, Special Issue on Computational Geometry'*, The Visual Computer, Vol. 1, 1985, pp. 69-97

[9] Inselberg A., Dimsdale B.: *'Parallel Coordinates: A Tool for Visualizing Multi-Dimensional Geometry'*, Visualization '90, San Francisco, CA, 1990, pp. 361-370

[10] Chernoff H.: *'The Use of Faces to Represent Points in k-Dimensional Space Graphically'*, Journal Amer. Statistical Association, Vol. 68, pp. 361-368

[11] Tufte E. R.: *'The Visual Display of Quantitative Information'*, Graphics Press, Cheshire, CT, 1983

[12] Pickett R. M.: *'Visual Analyses of Texture in the Detection and Recognition of Objects'*, in: Picture Processing and Psycho-Pictorics, Lipkin B. S., Rosenfeld A. (eds.), Academic Press, New York, 1970

[13] Pickett R. M., Grinstein G. G.: *'Iconographic Displays for Visualizing Multidimensional Data'*, Proc. IEEE Conf. on Systems, Man and Cybernetics, IEEE Press, Piscataway, NJ, 1988, pp. 514-519

[14] Beddow J.: *'Shape Coding of Multidimensional Data on a Mircocomputer Display'*, Visualization '90, San Francisco, CA, 1990, pp. 238-246

[15] Levkowitz H.: *'Color icons: Merging color and texture perception for integrated visualization of multiple parameters'*, In Visualization '91, San Diego, CA, October 22-25 1991

[16] Keim D. A., Kriegel H.-P.: *'VisDB: Database Exploration using Multidimensional Visualization'*, Computer Graphics & Applications, Sept. 1994, pp. 40-49

[17] Keim D. A., Kriegel H.-P., Ankerst M.: *'Recursive Pattern: A Technique for Visualizing Very Large Amounts of Data'*, Proc. Visualization '95, Atlanta, GA, 1995, pp. 279-286

[18] Ankerst M., Keim D. A., Kriegel H.-P.: "'Circle Segments': A Technique for Visually Exploring Large Multidimensional Data Sets", Proc. Visualization'96, Hot Topic Session, San Francisco, CA, 1996

[19] Keim D. A., Kriegel H.-P.: *'VisDB: Database Exploration using Multidimensional Visualization'*, Computer Graphics & Applications, Sept. 1994, pp. 40-49

[20] LeBlanc J., Ward M. O., Wittels N.: *'Exploring N-Dimensional Databases'*, Visualization '90, San Francisco, CA, 1990, pp. 230-239

[21] Feiner S., Beshers C.: *'World within World: Metaphors for Exploring n-dimensional Virtual Worlds'*, Proc. UIST, 1990, pp. 76-83

[22] Shneiderman B.: *'Tree Visualization with Treemaps: A 2D Space-Filling Approach'*, ACM Transactions on Graphics, Vol. 11, No. 1, pp. 92-99, 1992

[23] Johnson B.: *'Visualizing Hierarchical and Categorical Data'*, Ph.D. Thesis, Department of Computer Science, University of Maryland, 1993

[24] Robertson G. G., Mackinlay J. D., Card S. K.: *'Cone Trees: Animated 3D Visualizations of Hierarchical Information'*, Proc. Human Factors in Computing Systems CHI '91 Conf., New Orleans, LA, 1991, pp. 189-194

[25] Rekimoto J., Green M.: *'The Information Cube: Using Transparency in 3d Information Visualization'*, Proc. 3rd Annual Workshop on Information Technologies & Systems (WITS '93), 1993, pp. 125-132

[26] Battista G. D., Eades P., Tamassia R., Tollis I.: *'Annotated Bibliography on Graph Drawing Algorithms'*, Computational Geometry: Theory and Applications, Vol. 4, 1994, pp. 235-282

[27] Cleveland W.S.: Visualizing Data, Hobart Press, Summit, New Jersey, 1993

[28] http://www.ics.uci.edu/~mlearn/databases/

[29] Bertin J.: Semiology of graphics, W. J. Berg, Translated from Sémiologie graphique. Editions Gauthier-Villars, Paris, 1967, Madison, WI: The University of Wisconsin Press, 1983

[30] Olsen, K. A., Korfhage, R. R., Sochats, K. M., Spring, M. B. and Williams, J. G. Visualisation of a Document Collection: The VIBE System, Information Processing and Management, Vol. 29, No. 1, pp. 69-81, Pergamon Press Ltd, 1993

[31] Hoffman, P., Grinstein G., Marx, K., Grosse, I., Stanley, E., "DNA Visual and Analytic Data Mining", IEEE Visualization '97 Proceedings, pages 437- 441, Phoenix, AZ, 1997 [http://www.cs.uml.edu/~phoffman/viz]