DriftWeed – A visual metaphor for interactive analysis of multivariate data

Stuart Rose[1, 2], Pak Chung Wong[2]
stuart.rose@acm.org, pak.wong@pnl.gov

[1]Department of Mgmt. Info. Systems
University of Arizona
Tucson, AZ 85721

[2]Pacific Northwest National Laboratory
P.O. Box 999,
Richland, WA 99352

Abstract

We present a visualization technique that allows a user to identify and detect patterns and structures within a multivariate data set. Our research builds on previous efforts to represent multivariate data in a two-dimensional information display through the use of icon plots. Although the icon plot work done by Pickett and Grinstein is similar to our approach, we improve on their efforts in several ways.

Our technique allows analysis of a time series without using animation; promotes visual differentiation of information clusters based on measures of variance; and facilitates exploration through direct manipulation of geometry based on scales of variance.

Our goal is to provide a visualization that implicitly conveys the degree to which an element's ordered collection (pattern) of attributes varies from the prevailing pattern of attributes for other elements in the collection. We apply this technique to multivariate abstract data and use it to locate exceptional elements in a data set and divisions among clusters.

Keywords: Multivariate Visualization, Data Exploration, Data Mining.

1. Introduction

The main goal of icon plots is to produce a visualization that conveys information through stimulation of a person's pre-attentive visual processing capabilities [2,3]. Generally, an icon plot refers to a display consisting of multiple icons (glyphs) in which the focus is on a shifting perceptual pattern rather than the individual features of each glyph. A glyph is a visual icon consisting of one or more visual features whose characteristics are controlled by values associated with the represented information element. A glyph that visually reflects data values of the associated element is referred to as being data-controlled [1,2,5]. A shift in the value of the data results in a shift of its associated visual feature.

The Pacific Northwest National Laboratory is managed for the U.S. Department of Energy by Battelle Memorial Institute under Contract DE-AC06-76R1-1830.

Icon plots focus on graphically encoding features of information elements so that the human eye will naturally pick up on those features with significant characteristics. Pickett and Grinstein's icon plots develop shifting textures where a boundary between texture regions identifies a shift in the characteristics of the represented group of elements [2]. This focus on pre-attentive processing is advantageous because it eliminates the need for a look-up table to decode each glyph. By drawing attention to shifts in the visual field rather than the actual features of individual glyphs, a user is able to quickly assess where shifts in characteristics occur within a large collection of elements.

An effective icon plot builds a representative glyph for each information element so that differences among glyphs reflect differences among elements. An advantage of icon plots is that shifts in texture are often improved by increasing the amount of overlap between glyphs in the icon plot. The icon plot is then able to represent a larger number of elements than would be possible without overlapping glyphs.

To promote sensible boundaries in the visualization where real shifts in features occur, a sensible ordering/placement scheme should be employed. Typically, labeled axes that are associated with a continuous measure are used, such as a map of a physical location [2,4] or as in [8] where one axis represents the age of the victim and the other axis the age of the felon. Other placement schemes can be adapted from Self-Organizing Maps, multidimensional scaling (MDS), or a bin scheme such as that used in the SunflowerVisual Metaphor [6]. To promote pre-attentive processing, the placement of glyphs within the plot should be at regular intervals with few gaps, a requirement that has probably prevented its application to MDS due to MDS' focus on clustering rather than equal spacing. Time series information also lends itself well to the application of icon plots and is the ordering scheme we have used to test our system.

2. Method

We define a new method for constructing a data-controlled glyph that will be represented within an icon plot. We provide a variable scaling function that controls the visual manifestation of components within instances of the new glyph. We also provide a tool that gives the user control over the degree to which scales of variance are visually differentiable from one another in the icon plot.

2.1 The Glyph

Each information element in the collection is displayed as a glyph in the icon plot. Each glyph consists of an ordered series of connected feature segments. The geometric character of each feature segment is determined by two measures corresponding to that segment's associated feature. One measure controls the horizontal component of the feature segment; the other measure controls the vertical component. In the current implementation, each measure reflects variance associated with the information element's feature and the pattern of features occurring within the collection. As a result, deviations in feature patterns can be represented and detected in the display.

The order in which feature segments are located within the glyph may be identical to the order of attributes in the entity that the glyph represents, or the user may redefine the order. Taken together, the ordered series of connected feature segments represents the features of an element. The order of feature segments may match the order in which those attributes occur or may be reordered in a manner more appropriate to the visualization. Glyphs are then placed along an axis in a meaningful order.

In our first implementation, we represent each of the feature segments as line segments. The origin of each consecutive line segment in a glyph is located at the unattached endpoint of the previous line segment (see Figure 1). The use of line segments as feature segments creates a visualization in which a gradual drift is noticeable for most of the glyphs. Although this often has advantages because groups of similarly featured glyphs will drift in the same direction, in some instances it may be desirable to eliminate this drift.

To reduce the amount of drift that a glyph exhibits, an image can be used for the feature segment. Each feature segment then consists of a small image that is scaled along the horizontal and vertical axes according to its associated measures, as shown in Figure 3. The glyph then follows a distinctly vertical orientation, as shown in Figure 3.

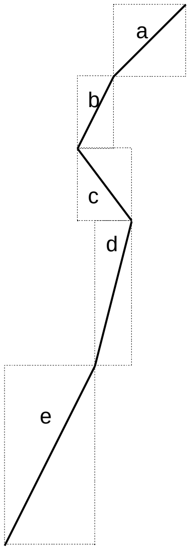

Figure 1

In determining what image to use for the feature segments, it helps to both maintain continuity among adjacent feature segments and to provide a distinguishing feature. We construct an image whose graphic consists of a doubleback line, such that the glyph's continuity is maintained between adjacent feature segments. Manipulating the horizontal and vertical components stretches this image but does not destroy the continuity of the overall line created, as the endpoints match up between feature segments. Figure 4 shows an additional image that may represent feature segments and its corresponding line plot.

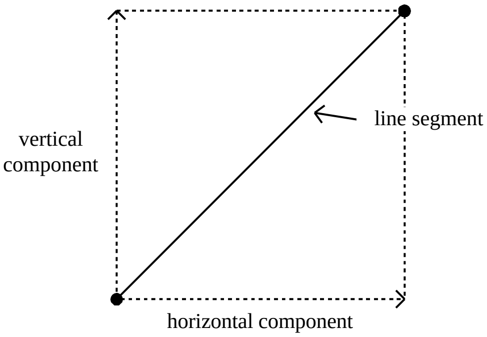

Figure 2

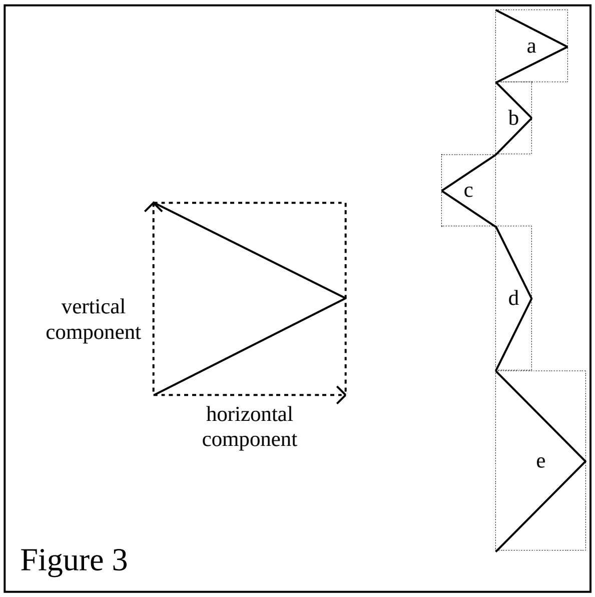

Figure 3

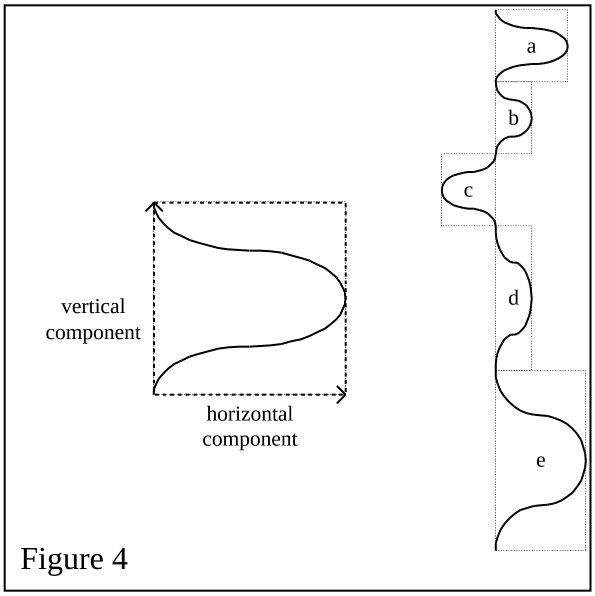

Figure 4

Although not investigated in our prototype, the line plots can be further manipulated to convey another variable or dimension by adjusting parameters for each feature segment such as roundness, slant, polar rotation, color, hue, opacity and shape of the image's contained graphic.

2.2 Variable scale

The information encoded within each element's glyph provides insight into several aspects of the element in such a way that these aspects of the glyph may periodically occlude one another in the display in such a way as to obscure meaningful aspects of the visualization. Each feature segment represents two measures of a particular feature. A user may be more interested in one of the two measures or may be interested in reducing the impact that one of the two measures has on the display.

We provide a variable scale transformation interface device, three examples of which are shown in Figure 5, that gives the user control over the visible magnitude of each feature segment's component based on its associated measures. The variable scale transformations let the user focus on specific components and attributes of each element. This adjusts the relative measures of variance associated with each feature segment. The scale also determines the magnitude of each feature segment, given its associated measure of variance. The horizontal and vertical components are controlled independently.

Giving the user direct control over the impact that variance between attributes has on the visualization promotes the identification of distinctions between elements.

Each of the grid lines in Figure 5 represents the relative magnitude associated with a measure of variance. The scrollbars are used to adjust constants that are applied to the components to determine the size with which a feature segment is drawn to the display. This enhances the visible difference among elements' patterns of

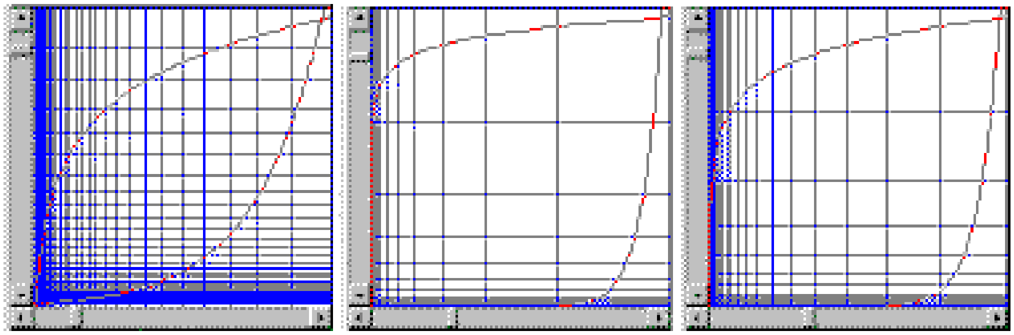

Figure 5

attributes and aids the user when visually isolating exceptional elements or patterns in the collection.

A line segment's contribution to an icon plot's height, span, or drift is therefore dependent on the variance associated with the attribute it represents and the selected magnitude that its measure of variance is assigned in the visualization.

2.3 Managing occlusion

It is often desirable to represent as many elements from a collection as is possible for the given display space. A consequence of our interest in the variance that a glyph exhibits from its counterparts is that we are able to place glyphs much closer to one another than if we needed to identify individual components of a glyph. We take advantage of this focus on variance by stacking glyphs on top of one another in the display. In order to draw 3000 glyphs in a space only 600 pixels wide, we place several at each distinct location along the horizontal axis.

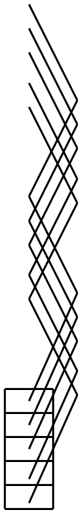

Figure 6

The origin of each element's glyph is incremented along the vertical axis as the number of glyphs at that horizontal position increases. This reduces the overlap that the glyphs would normally incur by being placed at the same location. As a result, if two glyphs at the same horizontal position are identical, they will not completely obscure one another. Furthermore, stacking of glyphs does not prevent our primary activity of recognizing differences among elements because the glyphs we are interested in will have markedly different geometric characteristics from their neighbors.

It should be noted that we have essentially added a third dimension to our icon plot. Because the visualization displays strong associations among elements at the same horizontal location instead of proximity, greater significance can be made through stacking.

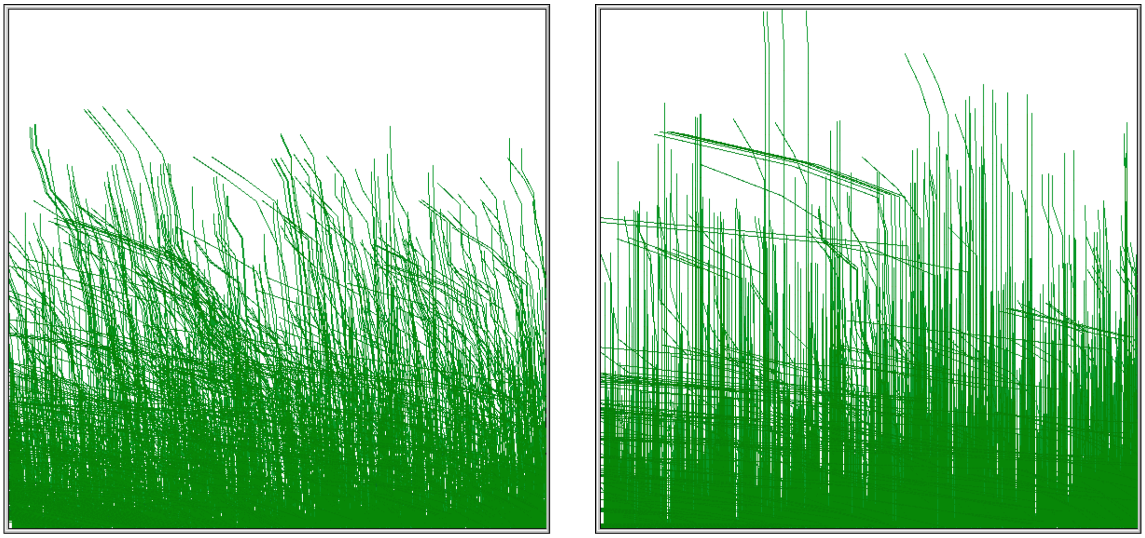

Figure 7

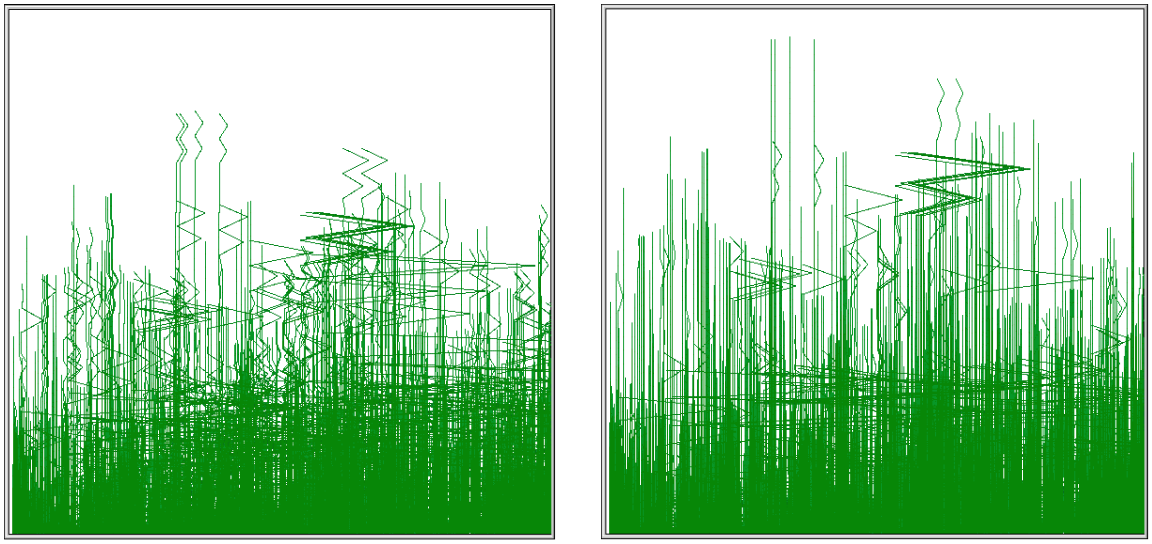

Figure 8

3. An Example

Figures 7 and 8 show output from the analysis of a dataset consisting of approximately 3200 news documents following the terrorist bombing in Oklahoma City. Topics occurring in the documents are considered features of the elements in this dataset. Statistical data on each of the documents is obtained and further analyzed within our prototype to calculate the variance associated with each of the features as well as the patterns of those features in the documents.

The vertical component associated with each feature segment reflects the infrequency of that feature's occurrence in the dataset. Tall line plots in figures 7 and 8 therefore represent documents that have infrequent, or exceptional topics. The horizontal component reflects the variance between the pattern of occurrence for the feature in the element and the predominant pattern of occurrence of that feature in the dataset. Noticeable horizontal deviations reveal documents whose combinations of topics are unlikely or exceptional, given the current dataset.

It is interesting to note that there are several similarly shaped line plots that appear two or more times within the visualization. It is highly likely that these similar line plots represent copies of the same document. We are therefore seeing a feature of the dataset in which news documents that have been collected are periodically repeated.

We have used a measure of variance based on statistical measures of occurrence to control the visual features of the line plots. It is also possible to apply other measures to the line plot, such as similarity to a user-specified query so that the most distinguishable line plots will represent documents containing features of interest to the user.

4. Conclusion

We have presented a new visualization technique that may be used to visually segregate clusters of information elements based on degrees of variance with the prevailing characteristics of the collection. A hierarchy of variance is presented in the display, which the user can adjust and control to discover clusters, patterns, and exceptions within the collection. When used to identify exceptions in a collection, the number of elements that can be shown is limited not by the size of the display but by the processing capability of the machine.

Our basic construct for a new icon plot is useful for the identification of exceptions within a collection of data. In an effort to show the potential of the basic concept, we have restricted the implementation to one color. Future implementations may be improved by manipulating a color range or transparency in order to draw greater attention to salient features within the visualization. It is important to be able to manipulate the relative effect that scales of variance have on their visual counterparts. While the horizontal and vertical components of each line segment do interfere with each other at times, we feel that in most cases the two complement each other and that any conflict can be mediated through manipulation of the scaling interface associated with either component.

5. Acknowledgement

This research was supported by a Laboratory Directed Research and Development grant funded by the U.S. Department of Energy for the Pacific Northwest National Laboratory. Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle Memorial Institute under contract DE-AC06-76RLO 1830. The first author was a graduate fellow at PNNL where the research was conducted over the summer. We wish to thank Kris Cook, Sharon Eaton, Dennis McQuerry, Rik Littlefield, and Jim Thomas who provided assistance of many forms throughout this research.

6. References

1. P. C. Wong and R. D. Bergeron. 30 Years of Multidimensional Multivariate Visualization. *Scientific Visualization: overviews, methodologies, and techniques,* Gregory M. Nielson, Hans Hagen, and Heinrich Muller, editors, IEEE Computer Society Press, 1997, pp. 3-34.
2. R. M. Pickett and G. G. Grinstein. Iconographic Displays for Visualizing Multidimensional Data. *Proceedings of IEEE Conference on Systems, Man, and Cybernetics'88,* Piscataway, NJ, 1988, pp. 361-370.
3. R. Rose, P1000 Science and Technology Strategy for Information Visualization. P1000 Visualization Planning Committee, 16 September 1996, version 2.
4. R. Erbacher, D. Gonthier, and H. Levkowitz. The Color Icon: A New Design and a Parallel Implementation. *Proceedings of the SPIE '95 Conference on Visual Data Exploration and Analysis II*, San Jose, CA, February, 1995, pp. 302-312.
5. E, Hamori and J. Ruskin. H Curves, A Novel Method of Representation of Nucleotide Series Especially Suited for Long DNA Sequences, *The Journal of Biological Chemistry*, Vol. 258, No. 2, pp. 1318 – 1327, 1983.
6. S. Rose, The Sunflower Visual Metaphor: A New Paradigm for Dimensional Compression. *IEEE Proceedings of the Symposium on Information Visualization*, pp. 131-134, 1999.
7. Arning, R. Agrawal, and P. Raghavan. A Linear Method for Deviation Detection in Large Databases. *Proc. of the 2nd Int'l Conference on Knowledge Discovery in Databases and Data Mining*, Portland, Oregon, August, 1996.
8. http://www.cs.uml.edu/~grinstei/fbi.jpg