Dual View: Multivariate Visualization Using Linked Layouts of Objects and Dimensions

David Borland*
RENCI
University of North Carolina at Chapel Hill

David Gotz†
SILS
University of North Carolina at Chapel Hill

ABSTRACT

The display of multivariate data is a common task in data visualization. However as dimensionality increases, it becomes increasingly difficult to visualize all dimensions using standard multivariate visualization techniques, such as parallel coordinates. Dimension reduction is often used to show relationships between data objects in a lower-dimensional representation, but the relationships between data objects and the original dimensions is typically lost. We introduce Dual View, a visualization technique for high-dimensional datasets that directly represents both data objects and data dimensions in separate 2D layouts. Linked views, spatial aggregation, and iterative layout refinement enables the exploration of high-dimensional datasets. We present the underlying algorithms for layout and interaction, a prototype Dual View user interface, and some examples applying Dual View to multidimensional datasets.

Index Terms: H.5.2 [Information Systems]: Information Interfaces and Presentation—User Interfaces; I.3.8 [Computing Methodologies]: Computer Graphics—Applications

1 INTRODUCTION

The collection and analysis of very large and complex datasets has become widespread across a range of domains in recent years. Data visualization is often employed to help analyze and understand such datasets, and many existing visualization techniques can be directly applied in cases of increased data *volume*–e.g. a bar chart showing the distribution of a binary variable works equally well for ten records as for 1 billion records. However increasing data *complexity*, including many dimensions of different types, can be problematic for common visualization techniques. Complex high-dimensional datasets can contain hundreds, or thousands, of variables. In contrast, the most complex example in a survey of the state of the art in parallel coordinates contains just eight dimensions [3], and a related technique includes examples of up to 26 visualized dimensions [1]. Dimension reduction techniques such as principal component analysis (PCA) and multidimensional scaling (MDS), are often used to visualize multidimensional datasets by showing relationships between data objects in a lower-dimensional representation. However, these techniques typically fail to retain the relationships between data objects and the original dimensions.

Dual View is a visualization technique designed to address this challenge. It directly represents both data objects and data dimensions in separate 2D layouts (Figure 1). Dual View combines linked views of relationships among and between objects and dimensions, spatial aggregation, and iterative layout refinement to enable the exploration of high-dimensional datasets. Moreover, it can be applied to both numeric and categorical dimensions. We present the underlying algorithms, a prototype Dual View user interface, and some examples applying Dual View to real world datasets.

*e-mail: borland@renci.org
†e-mail: gotz@unc.edu

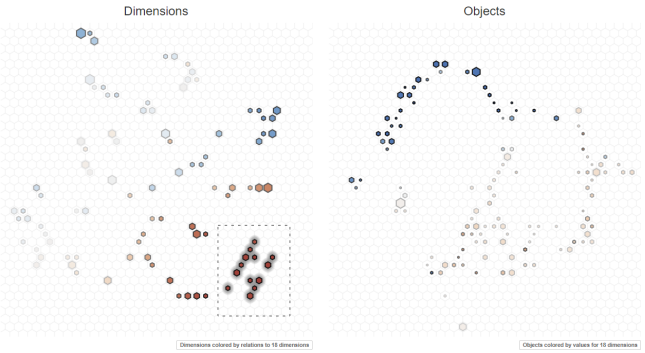

Figure 1: A Dual View visualization of a multivariate dataset with 167 dimensions (molecular shape descriptors) on the left, and 476 objects (molecules) on the right. Interactive highlighting of 13 cells representing 18 similar dimensions (dashed box) shows the relation of these dimensions to the other dimensions and data objects via color. For example, a group of objects with similar low values for those dimensions can be seen in blue.

2 METHODS

The fundamental concept of Dual View is to provide a 2D view of data objects, as is common with dimension reduction techniques, along with a separate 2D view of the data dimensions, such that each dimension and each object is directly represented.

2.1 Relations

Relations calculated (a) among dimensions, (b) among objects, and (c) between objects and dimensions, are used for 2D layout and interactive highlighting.

Dimension Relations. For dimension relations, different methods are used depending on the dimension types. *Numeric-numeric* relations are computed using Pearson's r, which results in a value in the range [-1, 1], where -1 is total negative correlation, 0 is no correlation, and 1 is total positive correlation. *Numeric-categorical* relations are computed using multiple regression with $k-1$ variables for a categorical variable with k labels. The regression R^2 value provides a value in the range [0, 1], with 0 indicating no association, and 1 indicating that 100% of the variability in the numeric variable is explained by the categorical variable. *Categorical-categorical* relations are computed using Cramér's V, which also provides an association value in the range [0, 1].

Object Relations. For relations between objects, the object similarity across all dimensions is computed. Our prototype uses cosine similarity, though other similarity metrics could also be used.

Dimension-Object Relations. Relations between dimensions and objects are determined by a normalized value, v' (in the range [0, 1]) from the object value v for that dimension d. For numeric dimensions, this is a straightforward linear mapping: $v' = \frac{v - d_{min}}{d_{max} - d_{min}}$. For categorical dimensions with k labels there is no inherent mapping from a label a number in the range [0, 1]. In order to emphasize

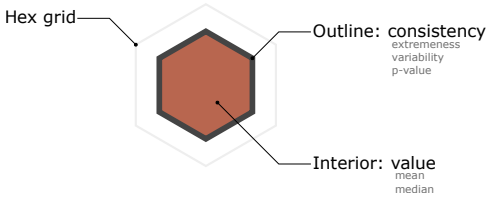

Figure 2: Overview of hexagonal glyph.

"rare" values, each label is sorted by decreasing frequency, and then mapped to [0, 1] based on the sorted index v_i: $v' = \frac{v_i}{k-1}$.

When mapping to color, dimension relations map values from [-1, 0, 1] to [blue, light grey, red] in order to be able to display negative vs. positive correlations. Categorical relations (in the range [0, 1]) will this always range from light grey to red. All other relations map values from [0, 0.5, 1] to [blue, light grey, red].

2.2 Layout

We use t-SNE [4] for our 2D layouts to emphasize clusters in the data, although other techniques (e.g., PCA, MDS) could also be used. Dimensions layouts use the computed dimension relations. For numeric-numeric dimension relations, the absolute value of the correlation, $|r|$, is used such that highly related dimensions should be clustered together, even if the relation is negative. Objects are laid out using the normalized values for each dimension. The user can control t-SNE parameters such as perplexity, early exaggeration, learning rate, and number of iterations. In order to handle over-plotting issues when rendering many data points, hexagonal binning is used [2]. Each hexagonal grid cell containing at least one data point contains a hexagonal glyph with area proportional to the number of data points within the cell (Figure 2), normalized to the maximum number per grid cell in each layout.

2.3 Interaction and Visualization

Users can select dimensions and objects via clicking individual grid cells, or dragging a selection rectangle. Selected cells are represented with a grey halo. For a given selection, a selection *connection* is computed for each data object and dimension, representing the distribution of the relations between that dimension/object and the selection. Each connection consists of two components: a central tendency *value* and a *consistency* measure. The interior of the hexagonal glyph is colored as described in Section 2.1, and the consistency is displayed via the darkness of the glyph outline. The current prototype enables the user to choose between mean or median for the value, and standard deviation, p-value, or extremeness for consistency. For a selection size n, the extremeness is calculated as the average normalized distance from the midpoint of the normalized range for that relation, e.g.:

$$\frac{1}{n}\sum_{i=1}^{n}\left|(v_i'-0.5)*2\right|$$

The extremeness helps differentiate between (a) connections with consistent mid-range relations to the current selection, and (b) connections with distributions of high and low relations (Figure 3). Since each hexagonal glyph can represent multiple data dimensions or objects, the average value and consistency are displayed.

Since both dimensions and objects can be selected at the same time, the dimension and object views both adopt priority-based highlighting. For the dimension layout, selected objects take precedence, and vice versa. The user can also mouse-over any populated cell to see a tooltip with the dimension/object names in that cell. In the absence of a selection, this cell will be used for highlighting. Finally, the user can choose to recompute the t-SNE layout for dimensions

Figure 3: Hexagonal glyph appearance for different selection connection distributions, mapping extremeness to outline intensity.

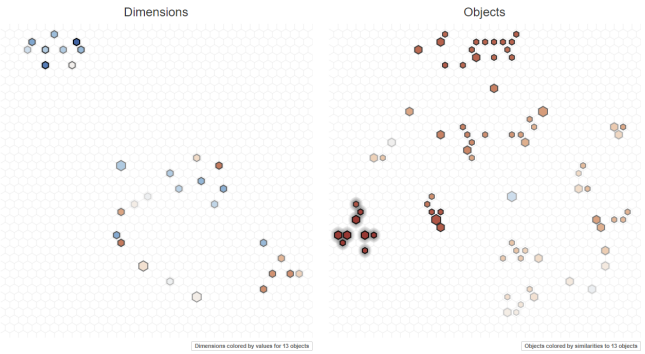

Figure 4: A visualization of a multivariate dataset in which the user has refined the dimension layout based on the highlighted objects. A cluster of similar dimensions (various automobile attributes) with respect to those objects (automobiles) can be seen in the top-left of the dimension layout.

using just the selected objects, and vice versa, enabling the user to iteratively explore alternative layouts based on different combinations of objects and dimensions (Figure 4).

3 Limitations and Future Work

The Dual View technique is intended to provide a useful overview of large multidimensional datasets to identify potential relationship of interest for investigation. Future work will include incorporating application-specific supplementary visualizations to provide detailed views of the user's analytic focus, as well as expanding the interactive capabilities to include concepts such as multiple groupings of objects and dimensions.

Although the visualization design is designed to handle arbitrary numbers of dimensions and objects via aggregation, current performance limits its applicability to a few hundred dimensions/objects, as all computation is performed on the fly in the browser. We plan on exploring ways to increase the performance, including precomputation of relations.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1704018.

References

[1] D. Borland, W. Vivian L., and W. E. Hammond. Multivariate visualization of system-wide National Health Service data using radial coordinates. In *Proceedings of the 2014 Workshop on Visual Analytics in Healthcare (VAHC 2014)*, 2014.

[2] D. B. Carr, R. J. Littlefield, W. L. Nicholson, and J. S. Littlefield. Scatterplot matrix techniques for large N. *Journal of the American Statistical Association*, 82(398):424–436, 1987. doi: 10.2307/2289444

[3] J. Heinrich and D. Weiskopf. State of the Art of Parallel Coordinates. In M. Sbert and L. Szirmay-Kalos, eds., *Eurographics 2013 - State of the Art Reports*. The Eurographics Association, 2012. doi: 10.2312/conf/EG2013/stars/095-116

[4] L. van der Maaten and G. Hinton. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.