Visualization of High-Dimensional Clusters Using Nonlinear Magnification

T. Alan Keahey

Los Alamos National Laboratory
MS B287
Los Alamos, NM 87501

ABSTRACT

This paper describes a visualization system which has been used as part of a data-mining effort to detect fraud and abuse within state medicare programs. The data-mining process generates a set of N attributes for each medicare provider and beneficiary in the state; these attributes can be numeric, categorical, or derived from the scoring process of the data-mining routines. The attribute list can be considered as an N-dimensional space, which is subsequently partitioned into some fixed number of cluster partitions. The sparse nature of the clustered space provides room for the simultaneous visualization of more than 3 dimensions; examples in the paper will show 6-dimensional visualization. This ability to view higher dimensional data allows the data-mining researcher to compare the clustering effectiveness of the different attributes. Transparency based rendering is also used in conjunction with filtering techniques to provide selective rendering of only those data which are of greatest interest. Nonlinear magnification techniques are used to stretch the N-dimensional space to allow focus on one or more regions of interest while still allowing a view of the global context. The magnification can either be applied globally, or in a constrained fashion to expand individual clusters within the space.

Keywords: data visualization, nonlinear magnification, multivariate visualization, data mining

1. INTRODUCTION

This paper describes a system for multivariate visualization which has been designed with the particular needs of data-mining research in mind. The main goal of the system is to allow data-mining experts to better understand the results of the clustering algorithms which they use to identify "interesting" cases for further investigation. In this context, "interesting" cases are not simply the outliers, but are also found in the patterns of data, the outliers within clusters, and other more complex phenomena. The visualization system uses a number of tools to allow the data-mining researcher to better focus on items of higher interest, and is designed to work on larger problem sizes of the type that are often encountered in data-mining applications. The paper begins with a brief description of the fraud-detection application, then describes the basic system for multivariate visualization. After that, nonlinear magnification routines will be introduced for the selective magnification of clusters within the data, followed by a discussion of implementation issues and conclusions.

2. APPLICATION: MEDICARE FRAUD DETECTION

The cluster visualization system described in this paper was developed in 1996 as part of a data-mining project at Los Alamos National Laboratory, sponsored by the US Federal Health Care Finance Administration, for detecting fraud among Medicare providers and beneficiaries. The original data for examination is composed of N attributes (categoric, numeric or derived from the data-mining scoring process) for each data point (provider record). K-means analysis is then used to partition the N-dimensional space into 100 clusters, each represented by an N-dimensional cluster centroid. Each data point has a single cluster centroid associated with it, and each cluster centroid is associated with a number of data points are contained within the cluster. Associated with each of these centroids and data points is a *probability density function* (PDF) estimate which reflects the probability with which we would expect to find "normal" items within a given region of the N-dimensional space. For the specific examples in this paper, the dataset is composed of approximately $35,000$ 11-dimensional records for medicare providers from a single state in the southeastern United States.

Email: keahey@lanl.gov Web: www.cs.indiana.edu/hyplan/tkeahey/research/nlm/

3. CLUSTER VISUALIZATION

Many methods have been described for multivariate visualization, such as scatterplots,[1] dimensional stacking,[2] parallel coordinates[3] and virtual worlds.[4] In designing a visualization system for this particular application, we wanted a system which preserved as much of the spatial information from the original N-dimensional data set as possible. A major reason for wanting to preserve the spatial relationships is that much of the underlying data-mining process is based on spatial geometries and densities, and by keeping the visualization spatial in nature, we make for a much smaller cognitive step for the data-mining researcher who is using the visualization system to analyze his or her data-mining routines. A good discussion of these and other methods (which we sill not repeat here) can be found in.[5] The spatial nature of our system is most similar in flavour to virtual worlds,[4] although there are many significant differences.

3.1. Frames

As Feiner and Beshers pointed out in,[4] our human experiences with spatial positioning are inherently limited to 3D. As a result, while manipulation and visualization of 3D data are fairly natural tasks for us, it requires a large cognitive leap to extend these tasks to systems of higher dimensionality. Primarily for this reason, we describe our N-dimensional dataset in terms of $3D$ *frames* of the data. For a data set composed of N-dimensional points (each represented by an N-tuple $\{x_1, x_2, \ldots, x_n\}$) we can select *frames* of the data with each frame representing three of the N possible dimensions. Figure 1 shows one frame (three dimensions) of the eleven dimensions in our dataset. Each data record is rendered as a discrete point, and each cluster centroid is rendered as a wire-frame cube.

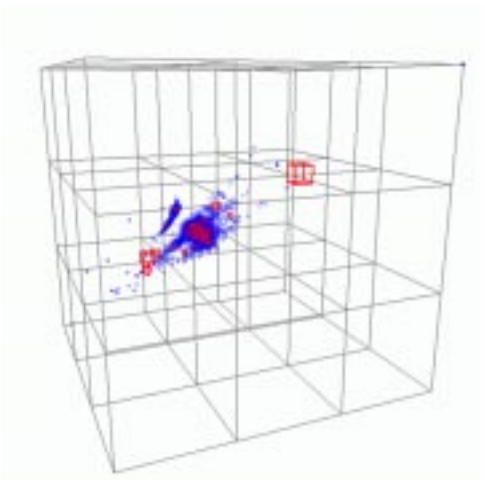

Figure 1. Visualization of 1 Frame (3 Dimensions)

Because of the relative sparsity of the cluster data, it is possible to effectively lay out more than three dimensions of the data in a single 3D coordinate system. Multiple frames can be laid out within the 3D coordinate space of the visualization, using colour cues to visually separate the dimensions. As an example, one frame could be composed of the dimensions $\{x_1, x_2, x_3\}$ and be rendered in green, while another frame could have the dimensions $\{x_5, x_4, x_8\}$ and be rendered in blue. Figure 2 shows a snapshot of the program with six dimensions of the data being rendered. As we will see in Section 3.3, brushing can be used to link data points between frames.

This layout of multiple 3D frames within a single 3D coordinate space is somewhat reminiscent of the virtual worlds within worlds of,[4] however there are significant differences. In our system, we do not constrain frames to a subregion of the encompassing space, instead each frame is scaled to fit fully into the encompassing space. While this

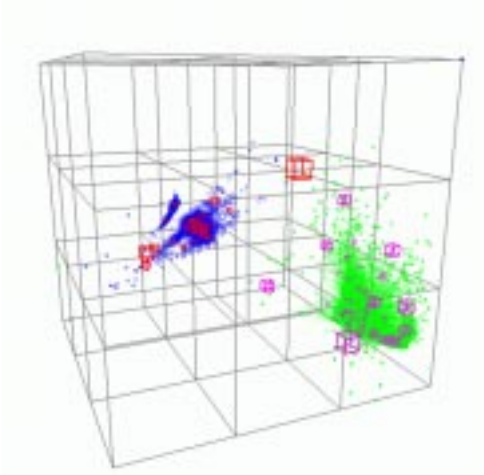

Figure 2. Visualization of 2 Frames (6 Dimensions)

may result in some intermingling of data between the frames, it also has the advantage of allowing for larger-scale views of the frames, and can increase the information density of the visualization by utilizing the empty spaces between clusters of data for the layout of other frames.

This method of overlaying frames within a single coordinate space would not necessarily be as effective for all multivariate visualization applications, however it has been noted that N-dimensional space is usually quite sparse,[5] so it seems likely that many other multivariate data sets could be amenable to this type of visualization. The clustering algorithms used in this data-mining process make the data space even more sparse than would normally be the case. The method used here brings the *small multiples* of Tufte[6] to mind, except that rather than creating a separate coordinate space for each set of dimensions we use a common space for all frames, thus resulting in *large multiples* which have the potential for making more effective use of the available space.

The method can be extended further to allow for three or more frames (9D+), however in practice we have only used it with one or two frames. The reason for this is that a primary goal in the design of this system was to provide the ability for the data-mining researcher to *compare* the effectiveness of different data-mining models along different dimensions. This comparison task is well matched to simultaneous visual comparison of two different frames. If additional frames were added for this task, it might complicate the comparison task unduly. It would be an interesting area for further research to investigate the usability of this system with different numbers of frames. It seems possible that since humans tend to hold no more than a handful of objects within short-term memory at any one time, that immediate comparisons between data points would become more difficult if the number of frames becomes too large.

3.2. Transparency

Individual records are rendered as single points, with a transparency that is inversely proportional to the PDF estimate so that "unusual" records are more clearly visible. Cluster centroids are rendered as wire-frame boxes, with box size being inversely proportional to the PDF estimate. Nonlinear scaling of the PDF values (similar to gamma-correction methods), can be used to interactively shift the transparency scale to focus on particular ranges of PDF values. Explicit clipping of items based on PDF values is also used to selectively render only those items having a given PDF value or lower. This allows the user to better focus on specific items of interest, and also reduces the size of the data set that must be rendered. Figure 3 shows how PDF clipping can be used on the dataset shown in Figure2 to reduce the total number of data points being rendered from about $35,000$ to about $2,000$.

3.3. Interaction

Selection of data records can either be done by entering the record ID number, or by selecting one or more data points by brushing with the mouse. When a record is selected, it is highlighted in all of the visible frames (i.e.

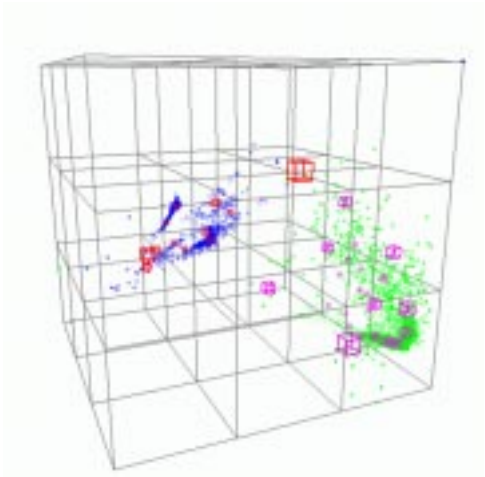

Figure 3. Using PDF Clipping to Isolate Interesting Cases

linked) to show its position relative to the other records in the visible dimensions. Once a record is selected, the user can highlight the cluster centroid associated with the record, and once a cluster centroid is selected, all of the data records associated with that cluster can be highlighted. Attribute dimensions can be dynamically assigned to frame dimensions, and the display smoothly interpolates between the source and target dimension assignments when this occurs, so that the user does not have to deal with sudden discontinuities. The user can rotate the space freely about the vertical axis, along with constrained rotation about the horizontal axis, so that views from any angle are possible without introducing the complexities of unconstrained 3D navigation. The user can also interactively zoom in on areas of interest, and magnification can be applied to any region or regions, as we will see in the next section.

4. NONLINEAR MAGNIFICATION IN 3D

The method for simulating a physical zoom in 3D described in the previous section allows for extreme close-up views of specific regions, but also introduces a number of problems in the process. The major problem with this type of physical zoom is that it is not possible to see both the details and the overall context; when you are zoomed in on a region most of the outlying space will not be visible, and when you are able to see the whole space the details are often too small as to be noticeable. Another characteristic of zoom systems is that it is only possible to focus on a single region at a time with them, which may present a problem for tasks where comparative visualization would require focusing on two or more regions simultaneously.

The term *nonlinear magnification* was introduced in[7] to describe the effects common to all of the many available approaches for stretching and distorting spaces to produce effective visualizations. The basic properties of nonlinear magnification are non-occluding in-place magnification which preserves a view of the global context. Most of the existing nonlinear magnification systems to-date have involved the magnification of 2D information spaces. Many of these systems also rely on perspective projections of a mapping of the 2D information space onto a 3D manifold in order to create the magnification effect,[8,9] and are thus constrained to the magnification of 2D spaces only. In contrast to these perspective-based systems, transformation-based techniques such as the nonlinear transformations of[7] and the hyperbolic spaces of[10] allow for simple and direct extension to 3 or more dimensions. This visualization system uses 3D versions of the transformations described in[7] simply by the addition of a z coordinate which is treated similarly to the x and y coordinates.

Visualization in three dimensions inherently involves occlusion of some portions of the data, however many methods for dealing with occlusion are available, such as clipping, transparency-based rendering, and creating "tunnels" through the data in order to see an occluded point of interest.[11] Occlusion does not present a great problem for this particular cluster visualization application however, as the data is composed of dense clusters in a relatively sparse space, and there is a great deal of empty space available in which to perform the magnification. Other likely candidates for 3D nonlinear magnification with similar sparsity properties might involve graph visualization as in.[10]

The techniques presented in[7] describe many different magnification effects that can be achieved using combinations of simple and computationally efficient 2D transformation functions. All of these effects have a straightforward extension to 3D viewing, and we will illustrate a few of them here. Figure 4 shows two examples of unconstrained magnification with a single center of magnification. The image on the left uses orthogonal magnification (independent magnification along x, y, z axes), and the image on the right uses radial magnification (magnification along line from center of magnification to the point). We perform the same nonlinear transformation as is performed on the data points on a regular grid covering the space; the transformed grid is then rendered in a unobtrusive colour and transparency in order to provide a visual cue to the user as to the overall pattern of magnification.

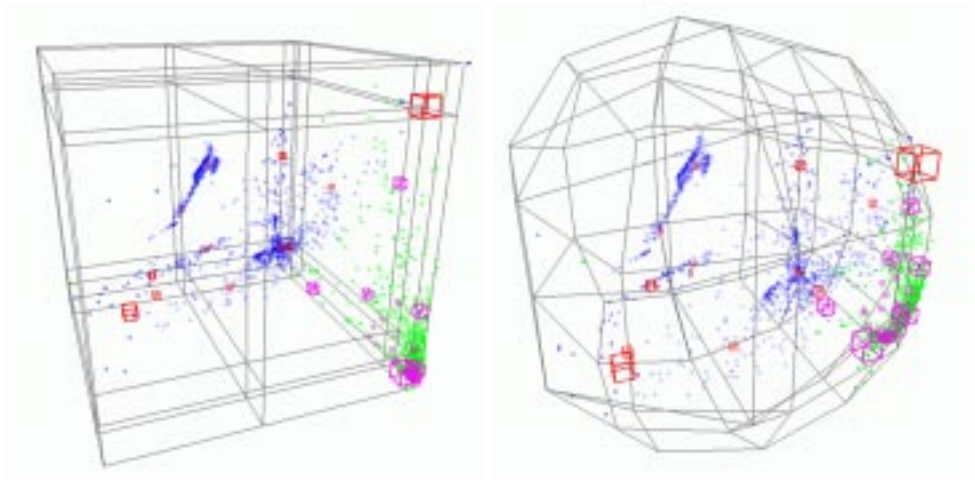

Figure 4. Orthogonal and Radial Unconstrained Magnification

As was also the case in,[7] there are several different ways in which we can combine multiple centers of magnification ("foci"). The example in Figure 5 shows the effect of averaging two unconstrained centers of magnification to produce a more complex magnification effect.

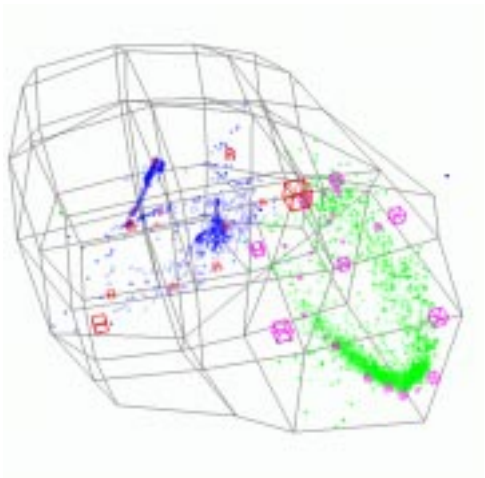

Figure 5. Combining Multiple Centers of Magnification

The previous examples involved the use of unbounded magnification that applied across the entire domain. It is also possible to put exact constraints on the domain that should be magnified, so that the magnification is localized. This provides a more static global context, which remains constant as individual portions of the space are magnified. Details for defining these constraints are provided in.[7,12] Figure 6 shows an example using constrained magnification to expand one set of clusters, while leaving the other clusters in the lower right-hand corner unchanged.

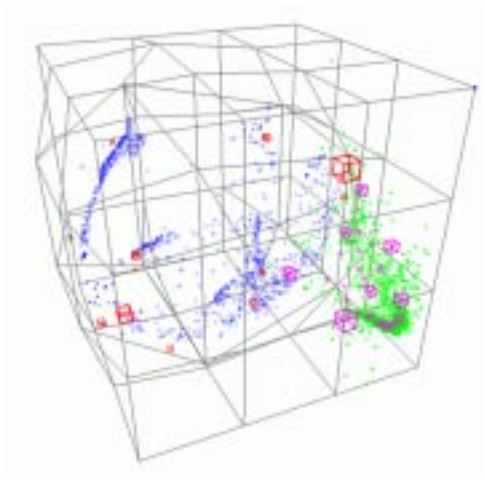

Figure 6. Constrained Magnification

We have already seen examples of combining multiple centers of magnification with unbound magnification. We can similarly combine multiple constrained foci to allow the user to simultaneously magnify specific clusters of data, as shown in Figure 7.

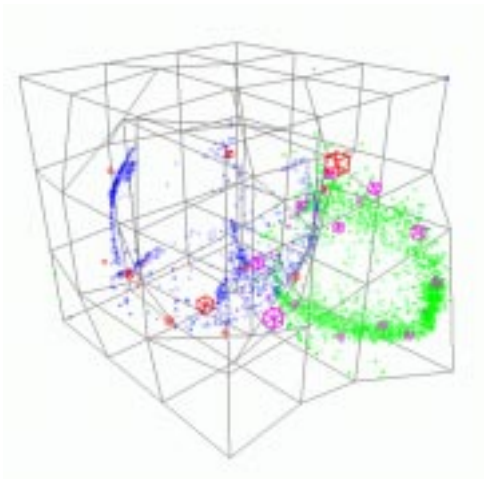

Figure 7. Constrained Magnification with Multiple Centers

There are a number of different ways in which the location of the centers of magnification can be manipulated. The centers can be moved independently along the x, y, z axes using the mouse, or they can be tied to selections of the data. In the latter option, when the user selects a specific cluster centroid, the center of magnification moves

via linear interpolation from the current location to the cluster location to provide a smooth animation resulting in a magnified target cluster area. Similar effects are possible with the selection of individual data points. Each center of magnification is independent of the others, so it is possible to leave some regions in a fixed state of magnification while exploring other clusters with a roving center of magnification.

5. IMPLEMENTATION

This visualization system depends on a high degree of interactivity to be effective. Because the data consists of discrete points, there is not a great deal of coherency inherent in it. Thus we provide simple tools (rotation, zoom, magnification) to allow the user to rapidly animate through a set of views to get a better feel for the nature of the dataset. It is crucial that these tools work in a smoothly interactive fashion for this to happen.

The nonlinear magnification routines used here are based on the FAD Toolkit originally described in.[7] Further information on this toolkit and availability can be found at: /www.cs.indiana.edu/hyplan/tkeahey/research/fad/ . Similarly to the H3 visualization system,[10] this system is designed to work on much larger problem sizes than are normally addressed by other nonlinear magnification-based visualization systems. The prototype system described here was designed to run at interactive speeds on a relatively slow SGI Indy workstation. For the examples with two frames of data, this amounts to transforming (magnifying) and rendering $70,000$ 3D points at near 10 frames per second. The toolkit library for producing the transformations is able to transform 3D points at a rate above 1.2 million points transformed per second (on a 250 MHz MIPS R10000), which would allow interactive rates on data sets of over $120,000$ points. The only nonlinear magnification based-system that we have seen which reports transformation rates within an order of magnitude of this is the H3 system,[10] although it is worth noting that the transformations provided here are of a more general nature and allow multiple interacting foci with constrained or global domains.

At this stage, the principle bottleneck to using this system with larger data sets lies in the rendering system. Much of the desktop hardware in use today is not capable of rendering $100,000$ 3D points at interactive speeds. Clipping based on PDF values is one way to reduce the size of the data set which must be rendered. In addition, the method used for generating transparency in the data points is also an issue. The fastest (and crudest) method uses polygon stipples to achieve screen-door transparency of the data points. The most expensive (and accurate) method uses alpha blending to simulate real transparency. Unfortunately the latter option requires back-to-front rendering of the data points, which implies a need for a full depth sorting of the points before rendering. A pseudo-transparency method can also be used which uses unsorted alpha-blending to provide some of the benefits of true alpha transparency without the high cost of sorting data points. The choice of which method to use can be switched at runtime by the user depending on the graphics power of the workstation being used.

6. CONCLUSIONS AND FURTHER WORK

This cluster-visualization application combines several different techniques to enhance visualization for data mining. The use of frames allows high-dimensional visualization, while the transparency based rendering helps to reduce visual clutter to focus on the more important items of interest. Nonlinear magnification is also employed to enhance the view of one or more clusters while simultaneously allowing a view of the global context.

There are several potential areas for further research and improvements to the prototype visualizer. One aspect that is missing from the current system is a visual representation of the cluster boundaries. Adding such a representation may help the data-mining researcher better understand the properties of the clustering algorithms. Implicit magnification fields were introduced in[13] as a method for determining the amount of magnification that is implicit in complex transformations of the type described here. By synchronizing rendering functions to this implicit magnification field, significant efficiency gains may become possible if data points below a certain magnification level are aggregated or eliminated from the rendering. This represents a specific instance of the *generalized detail-in-context problem* described in.[14] A method is also described in[13] that allows properties of the data to directly define spatial transformations as a field of scalar magnification values, these data-driven magnifications offer the potential to automatically provide visual enhancement of the regions which are of greatest interest to the user.

REFERENCES

1. M. A. Fisherkeller, J. H. Friedman, and J. W. Tukey, "PRIM-9: An interactive multidimensional data display and analysis system," in *Dynamic Graphics for Statistics*, W. S. Cleveland and M. E. McGill, eds., Wadsworth, Inc., 1988.
2. J. LeBlanc, M. O. Ward, and N. Wittels, "Exploring n-dimensional databases," in *Proceedings of IEEE Visualization*, 1990.
3. A. Inselberg and B. Dimsdale, "Parallel coordinates: A tool for visualizing multidimensional geometry," in *Proceedings of IEEE Visualization*, 1990.
4. S. Feiner and C. Beshers, "Worlds within worlds metaphors for exploring n-dimensional virtual worlds," in *Proceedings of the ACM Symposium on User Interface Software and Technology*, 1990.
5. M. O. Ward, "XmdvTool: Integrating multiple methods for visualizing multivariate data," in *Proceedings of IEEE Visualization*, 1994.
6. E. R. Tufte, *The Visual Display of Quantitative Information*, Graphic Press, 1983.
7. T. A. Keahey and E. L. Robertson, "Techniques for non-linear magnification transformations," in *Proceedings of the IEEE Symposium on Information Visualization, IEEE Visualization*, pp. 38–45, Oct. 1996.
8. J. Mackinlay, G. Robertson, and S. Card, "The perspective wall: Detail and context smoothly integrated," in *Proceedings of the ACM Conference on Computer Human Interaction*, pp. 173–179, 1991.
9. G. Robertson and J. D. Mackinlay, "The document lens," in *Proceedings of the ACM Symposium on User Interface Software and Technology*, pp. 101–108, 1993.
10. T. Munzner, "H3: Laying out large directed graphs in 3D hyperbolic space," in *Proceedings of the IEEE Symposium on Information Visualization, IEEE Visualization*, Oct. 1997.
11. M. Carpendale, D. Cowperthwaite, and F. Fracchia, "Distortion viewing techniques for 3D data," in *Proceedings of the IEEE Symposium on Information Visualization, IEEE Visualization*, pp. 46–53, 1996.
12. T. A. Keahey, *Nonlinear Magnification*. PhD thesis, Department of Computer Science, Indiana University, Dec. 1997.
13. T. A. Keahey and E. L. Robertson, "Nonlinear magnification fields," in *Proceedings of the IEEE Symposium on Information Visualization, IEEE Visualization*, Oct. 1997.
14. T. A. Keahey, "The generalized detail-in-context problem," in *Proceedings of the IEEE Symposium on Information Visualization, IEEE Visualization*, Oct. 1998.