Data Mining and Visualization: Some Strategies

Edward J. Wegman

Center for Computational Statistics
George Mason University
Fairfax, VA 22030-4444

1. Introduction

This paper is intended to be illustrative of some of the visualization tools we have exploited in a data-mining-like framework. Huber (1992, 1994) developed a taxonomy of data set size which characterized data by increasing orders of magnitude. In earlier discussions, Wegman (1995), I have pointed out that somewhere around $10^6$ to $10^7$ bytes appears to be the practical limit for visualization of data while massive data sets easily venture into the range of $10^{12}$ bytes. The human eye has approximately $10^7$ cones implying that visualizing one observation per cone would optimistically put the upper limit of visual resolution at about $10^7$ observations. (I refer the reader to Wegman (1995) for more details on the limits of visualization and of computational feasibility.) Thus, data mining as such cannot successfully exploit visualization for truly massive data sets without some modification of the raw data. In Wegman (1999) I have suggested several approaches including binning and thinning to reduce the size of data sets making visual analysis more feasible.

With this caveat made, I would like to illustrate in this paper how several more or less standard statistical tasks can be carried out visually. Our basic approach involves a combination of three tools, parallel coordinates multidimensional displays, the $d$-dimensional grand tour, and saturation brushing. In combination these three tools are available in a down-loadable software called ExplorN (available at ftp://www.galaxy.gmu.edu/pub/software/) and also a commercial version called CrystalVision soon to be available.

Parallel Coordinates is a multidimensional visualization tool discussed by Inselberg (1985) and employed for data visualization by Wegman (1990). In order to represent a $d$-dimensional point, the basic idea is to draw $d$ parallel axes labeling them according to the data variables. A point is then represented by locating the value of each variable (component) long its respective axis and then joining the resulting points by a broken line segment. Several such diagrams are found in this paper. A fuller discussion of the statistical and data analytic interpretations of parallel coordinate displays is given in Wegman (1990). In some sense a parallel coordinate display is a generalization of a two-dimensional scatterplot. We shall refer to “data clouds” even when strictly speaking the parallel coordinate display involves line segments.

The $d$-dimensional Grand Tour is a generalization of the 2-dimensional grand tour introduced by Asimov (1985). The basic idea of a grand tour is to look at a data cloud from all possible points of view. As implemented, the $d$-dimensional tour is a continuous geometric transformation of a $d$-dimensional coordinate systems such that all possible orientations of the coordinate axes are eventually achieved. The algorithm is described in Wegman (1991) and also in Wegman and Carr (1993). Coupled with the parallel coordinate display, these two techniques allow for an in-depth study of high dimensional data. Partial grand tours can be accomplished by holding one or more variables fixed. A grand tour is in some sense a generalization of a two-dimensional rotation although it is not a rotation in the conventional sense.

Saturation Brushing is a generalization of ordinary brushing. Ordinary brushing is accomplished by brushing a data cloud with a color for the purpose of visually isolating segments of the data. In data sets where there is considerable overplotting, ordinary brushing is potentially confusing in misleading, particularly where there is an animation such as rotation or grand tour. This is particularly the case because in many computer graphics algorithms, colors are drawn according to the z-depth, lowest z-depth points are drawn last. This can lead to apparently arbitrary changes of color and certainly gives no clue as to the amount of overplotting. In saturation
brushing, each point is assigned a highly desaturated color (nearly black) and when points are overplotted, their color saturations are added via the so-called $a$-channel. Thus heavily overplotted pixels have fully saturated colors whereas pixels with little overplotting remain nearly black. Saturation brushing is described by Wegman and Luo (1997) and is an effective method for dealing with large data sets. Coupled with parallel coordinates and the grand tour, these methods allow for an extremely effective visual approach to large, high-dimensional data.

2. Density Estimation and Rapid Data Editing

Saturation brushing can be used with either parallel coordinate displays or with ordinary scatterplots. In effect, by using saturation brushing with a single color (say white against a black background or gray against a white background) the brushed display becomes a density estimate. Local smoothing is not required except to the extent that the data is binned by the resolution of the screen. For ordinary high resolution displays this means there are approximately $1.3 \times 10^6$ bins or pixels. With the $a$-channel, this density estimation is accomplished with the same speed as the rendering of an ordinary two-dimensional image. Figure 2.1 shows the so-called Pollen data which consists of 3848 points. At first glance this data would simply seem to be five-dimensional with elliptical contours suggesting an uninteresting multivariate normal data set. However, when a desaturated parallel coordinate plot is examined in Figure 2.2 at least two interesting features can be observed. First a bright feature in the middle of the diagram suggests that there is a hidden structure. In addition the larger X-features in the display suggest that the overall elliptical clouds have a five-dimensional hole in the middle. The central structure can rapidly be isolated using a cutting tool. This is best done in the parallel coordinate display. The result is shown in Figure 2.3. The central structure shown in Figure 2.2 is actually 99 data points spelling the word EUREKA as illustrated in Figure 2.3. The five-dimensional hole mentioned above can be verified by doing a more conventional density estimate.

3. Inverse Regression and Tree-Structured Decision Aids

The combination of parallel coordinates, partial grand tour and saturation brushing can be used in conjunction to achieve a form of inverse regression and a sort of tree structured decision aid. The data we use here to illustrate is 8-dimensional bank demographic data containing approximately 12,000 observations. One of the variables is profit, a variable we will treat as the dependent variable. After appropriate data editing for unknowns, we brush the profit variable with either red or green. Those observations with negative profit (loss) are brushed red, those with positive profit are brushed green. In an additive color system, red and green together make yellow. Thus regions of the covariates that show up as yellow are neutral with respect profit or loss. However, regions that are predominantly red reveal demographics that cause the bank to lose money while regions that are predominantly green represent the demographics of desirable customers. The partial grand tour is used to animate the demographic variables. Because the $d$-grand tour as we have implemented it forms orthogonal linear combinations of the demographic variables, we can visualize linear combinations of the demographic variables as a function of the dependent variable (profit). Thus we have a tool for inverse regression. Moreover by isolating linear combinations of demographic variables that are predominantly green or red, we can create a tree-structured decision rule for distinguishing desirable customers from undesirable
customers. Unfortunately full color graphics are not available in the printed form of this paper so the full impact of this technique cannot be seen here. However, a full version is available with the color graphics at our website, URL www.galaxy.gmu.edu/papers/datamining&visualization.html. The reader is encouraged to examine the sequence of images on our website.

4. Variable Selection and Dimensionality Reduction

Parallel coordinate displays coupled with saturation brushing can allow us to select variables visually for the purpose of discriminant analysis. This is illustrated with the so-called SALAD data, a thirteen-dimensional data set containing spectral response of some 10,000 plus chemical samples. The data is shown in Figure 4.1. The thirteen variables are intensity of spectral response in thirteen color bands with increasing wavelength. The variable on the bottom axis is a classification variable and is brushed with one of three colors red, blue, or green according to the class of chemicals. The idea is to look for one or more of the spectral bands which adequately discriminates the three classes of chemicals. Although less apparent in the black and white version, the additive color feature can again be exploited. Red + blue = magenta, red + green = yellow, and blue + green = cyan. Variable B10 separates blue and red, and in fact shows two distinct red clusters. Unfortunately, B10 does not discriminate red from green. However, in parallel coordinate displays, the slope of the line segments matters considerably. The slope of the green line segments between B9 and B10 is substantially different from the slope of the red line segments between the same axes. Thus the variable B10-B9, a surrogate for slope, will distinguish red from green. Thus only two variables, B9 and B10 are adequate to discriminate all three classes and in fact also discriminate the two subclasses of the red. This dimensionality reduction allows for real-time discrimination of chemical agents. Again we encourage the reader to view the color images at our website.

5. Clustering and Classification

Our last example uses a combination of parallel coordinates and the grand tour to achieve a combination of clustering and classification. The data addressed here is a 12-dimensional dataset called the Oronsay sand data. This was treated more fully in Wilhelm et al. (1999). The basic idea is that samples of sand from various locations is taken and put through a series of sieves. The weight of sand isolated by each sieve is measured resulting in a distribution of particle sizes for each sand sample. The overall goal of the Oronsay sampling experiment was to classify Mesolithic sand samples as to whether they resembled contemporary beach sand or contemporary dune sand. Samples
from two different locations in the Scottish Hebrides were taken and are illustrated in Figure 5.1. Although not recognized in the original analysis done by archeologists, Figure 5.1. shows that the two locations have distinctly different particle size distributions. Thus only the location with larger sample size (the sample colored black) was used in our analysis. Figure 5.2 is the result of following the BRUSH-TOUR strategy. The basic idea is to brush all visible clusters in the original orientation of the data with distinct colors. Then allow a grand tour rotation until new clusters show up. Brush the new clusters and repeat the BRUSH-TOUR until no new clusters are apparent. One can be confident of having isolated all clusters when the clusters identified by this technique move coherently under the grand tour rotation. In Figure 5.2, the beach sand is brushed with green, the dune sand is brushed with red and the unknown Mesolithic sand is brushed with black. The conclusion of our experiment was that although the Mesolithic sand more closely resembled contemporary dune sand it was still distinctly different and was really in a cluster by itself. The complete Oronsay sand analysis is available at our website http://www.galaxy.gmu.edu/papers/oronsay.html.

Acknowledgements

The preparation of this paper was supported by the United States National Science Foundation with a Group Infrastructure Grant DMS-9631351 and by the United States Army Research Office under Grant DAAG55-98-1-0404, I

References


