A Visual Approach To Exploratory Data Mining

Russell Anderson, West Texas A&M University, USA
Musa Jafar, West Texas A&M University, USA

ABSTRACT

As the first step upon commencing an in-depth data mining analysis, students should become intimately acquainted with the data under study. In this paper, we present a methodology and set of custom tools that we have designed and developed for use in our data mining courses that allows students to efficiently and effectively accomplish this task. The tools create interactive visual presentations of the data, encouraging students to explore the data in search of patterns or relationships that would then be investigated in subsequent steps using sophisticated statistical and machine learning tools.

Keywords: Data mining, information visualization, visual analytics, parallel coordinate plots, scatter plots

INTRODUCTION

Data mining has been defined as "the process of discovering useful information in large data repositories" (Tan *et al.*, 2006). It focuses on the discovery of valid, previously unknown, and actionable patterns and relationships. In its search for patterns, associations and descriptions, it goes beyond the simple retrieval of base facts and aggregations. Data mining frequently employs complex statistical and machine learning algorithms to extract this information. Given the large volumes of data on which data mining analysts usually work, the question of where to begin is often asked. In this paper, we present an approach that we have designed for our Data Mining courses that uses three locally developed visual tools to teach and conduct some of the introductory steps of data mining. These tools complement the current out-of-the-box data mining tools (Microsoft, Oracle, IBM, SAS, etc.) typically used in a Data Mining course.

For our students, the first step in the initial analysis of a dataset is to assess the structure of the dataset itself (the metadata) rather than its contents. This includes the classes of objects represented, the characteristics of each attribute (data type, scale and unit of measure), and data organization (flat file, tree/hierarchical structure, network, etc.).

The next step is to conduct an exploratory review of the data. This is where our tools are employed. In conducting this analysis, we instruct students to seek answers to the questions listed below. To create this list, we began with one developed by Amar *et al.* (2005). We kept the tasks on their list pertaining to data analysis; eliminated processing tasks such as data sorting; then added tasks that we had gleaned from literature reviews of data mining and visualization research.

1. What are the characteristics of individual attribute values?
   a. For numeric data, what are the data ranges and possible values?
   b. For discrete attributes of either nominal, ordinal, or numeric type: what are the domain and cardinality of the attribute's set of values?
   c. How are the values distributed – uniform, normal, skewed, multi-modal?

2. What is the quality of the data?

a. Are there outliers? Outliers may be defined with respect to values along a single dimension or with respect to two or more dimensions.
b. Are there missing values within recorded observations?

3. Are there patterns or relationships between attributes or combinations of attributes?
   a. What is the nature of relationships between attributes – direct or inverse correlation, linear or non-linear?
   b. Is there interaction between attributes? That is, are relationships between two attributes influenced by the values of a third attribute?

4. How do the observations compare? Do all observations appear to originate from the same underlying population or do there appear to be distinct subsets of observations (clusters) within the dataset?

Note the distinction between questions 3 and 4. Question 3 is a column oriented assessment, while question 4 is row oriented.

Some of the above questions are best answered by computing, then reviewing, a set of summary statistics. Questions 1a, 1b, and 2b fall into this category. Yet we have found that for an initial exploration, visual presentations are superior for questions 1c (distribution), 2a (outliers), 3 (relationships), and 4 (clusterings).

Previous investigators have found that good visual representations invoke parallel processing within the brain (Card *et al.*, 1999; Shneiderman, 2002; Ware, 2004). The brain without conscious effort segments items within the image, then focuses attention on features of the image that are different with respect to color, shape, size, cardinality, and orientation. A well-designed visualization provides a "pop-out" effect focusing attention on answers to many of the initial analysis questions introduced above. Visual images are also easier to recall than tabular statistical presentations.

As described later in the paper, we use three visual presentations: scatter plots, parallel coordinate plots, and a correlation matrix. In the sections that follow, we first review the specific structure and applications of scatter and parallel coordinate plots with respect to the above analysis questions. Under the assumption that many of the readers are not as familiar with parallel coordinate plots, we present a more in depth look at the definition, characteristics, and research pertaining to such plots. Next, we present features of our tools that we implemented specifically to support the initial data exploration process. We conclude with a walk-through of a typical student exploration using two datasets: olive oil samples from Italy (Forina *et al.*, 1983) and remote earth sensing observations collected in Australia (Worcester Polytechnic Institute, 2005).

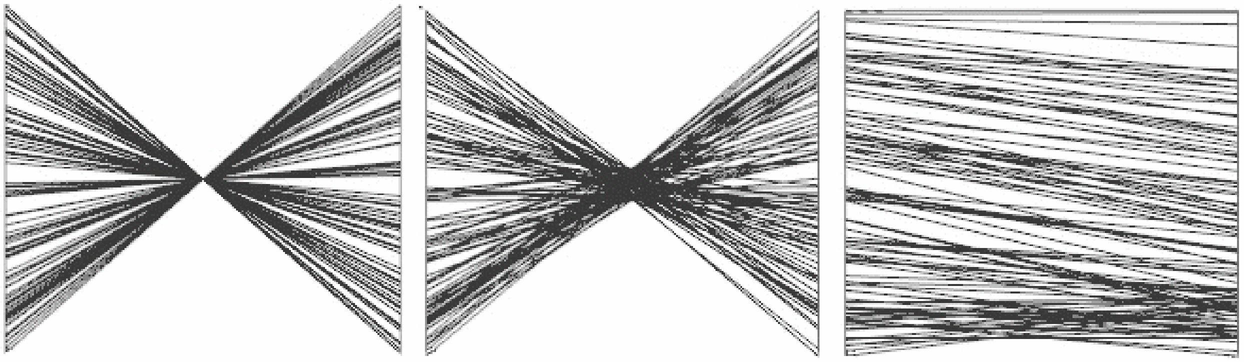
SCATTER AND PARALLEL PLOTS – DESCRIPTION AND APPLICATION

Scatter plots using orthogonal axes have been found to be highly effective for many of the initial analysis tasks – especially in assessing attribute relationships and distributions, and in outlier detection. They are especially natural in representing data with spatial dimensions or a temporal dimension in which the analyst attempts to assess changes over time. A scatter plot augmented with animation over the temporal dimension, has also been shown to effectively represent this component. They however, suffer from one major drawback, in that they only adequately support concurrent visualization of a limited number of dimensions. Scatter plots best depict the relationship between two continuous numeric attributes. It is, however, possible to effectively add a third continuous dimension using an XYZ axis plot. Scatter plots are not as effective when mapping discrete attribute values on the axes. Nominal attributes are best encoded by varying the color and/or shape of the scatter plot points.

To support the concurrent display of multiple dimensions, Inselberg (1985) first defined, then laid the theoretical foundation of parallel coordinate plots. Such plots are laid out as a sequence of potentially unlimited parallel axes – one for each attribute of interest in the dataset. Each observation in the dataset is represented by a set of connected line segments connecting the points on each axis corresponding to the attribute value for the observation.

There are many advantages to parallel plots, the most obvious being their ability to concurrently display more than three dimensions. With respect to the data analysis tasks listed previously, parallel coordinate plots, depending on the number of observations and number of dimensions can be most effective in assessing relationships between attributes, clustering of observations, finding outliers, and assessing distributions.

For example, with respect to assessing relationships, Inselberg (2002) , Mousafa and Wegman (2003) have pointed out that the connecting line segments between adjacent axes for two attributes that are inversely related will intersect. In a perfect inverse relationship, there will be a single point of intersection (Figure 1a), whereas a less than perfect inverse relationship will yield a broader spread of intersection points (Figure 1b). In a perfect direct linear relationship, the connecting line segments will be horizontal – given that the attribute values have been normalized. In a typical direct relationship, where there is a non-linear relationship and some residual error, the lines will mostly have a non-zero slope (indicating a non-linear relationship) while some will intersect (Figure 1c) as a result of errors in measurement or variances in the source population.



a) Correlation = -1.0 b) Correlation = -0.97 c) Non-linear with Residual Error

Figure 1: Bi-Variate Relationships

PLOT FEATURES FOR INTERACTIVE EXPLORATION

In this section, we review the features of our implementations of scatter plots, parallel coordinate plots, and the supporting color encoded correlation matrix. In all tool design decisions, we considered important the need for real-time student interaction with the plots in order to encourage exploration. We begin with the visual correlation matrix.

After a student loads the data set into the tool they are presented with a correlation matrix, visually presenting a pair-wise correlation of all attributes in the chosen dataset. See Figure 2.

Each cell within the matrix represents a color encoding of the correlation between the corresponding attributes. The more saturated the blue, the stronger the direct correlation and the more saturated the red, the stronger the inverse correlation. As the user hovers over a cell, the coefficient of correlation is displayed and the column and row labels for the given cell are highlighted. When a user clicks on a cell, a 2-d scatter plot of the corresponding cell attributes is opened for viewing. Figure 6 is an example of a 2-d scatter plot.

The scatter plot implementation creates plots of both 2and 3 orthogonal dimensions using either numeric, nominal, or ordinal attributes and a fourth dimension using color encoding of a nominal attribute. Students may interactively select which attributes to assign to each axis and to the nominal color attribute by selecting from list boxes in a control panel positioned to the side of the plot. See Figure 3.

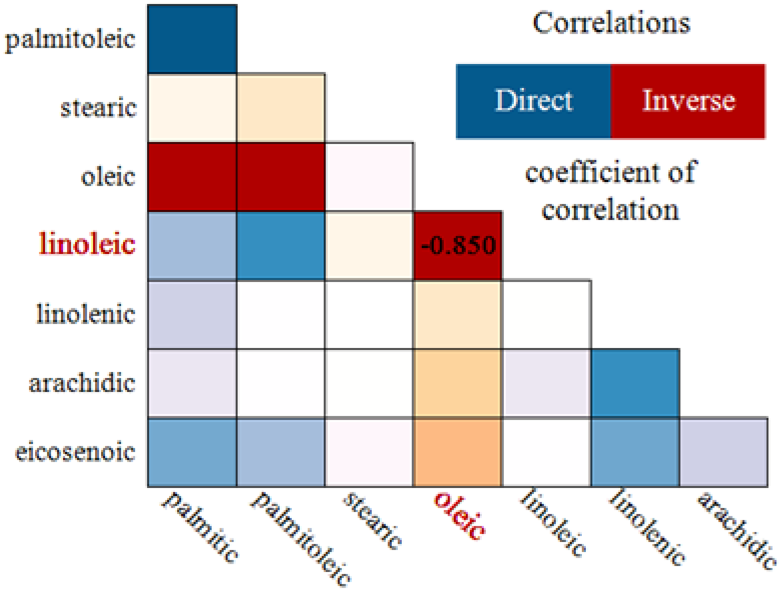
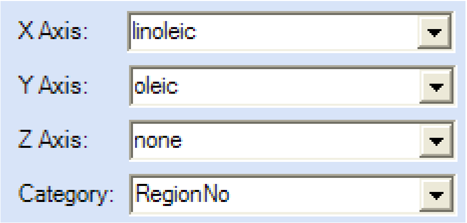


Figure 2: Correlation Matrix



X Axis: linoleic

Y Axis: oleic

Z Axis: none

Category: RegionNo

Figure 3: Axis Selection

Data points representing observation values between two continuous numeric attributes or between one numeric and one discrete attribute are drawn as small spheres of uniform size on the plot. Data points representing observations values between two discrete attributes are enlarged proportional to the number of observations at that location.

The visual presentations are rendered in three dimensions using the OpenGL graphics API. However, visual perception of a static three dimensional image on a two dimensional medium is difficult – especially one consisting mostly of lines or points. To overcome this problem, we allow the user to rotate the presentation about both the X and Y display axes by simply dragging the mouse pointer over the image in a desired rotation direction. As the image smoothly rotates, the brain readily assesses the 3-d structure of the image much more efficiently than is possible with static 2-d projections of a 3-d structure. Figure 8 is an example of a rotated 3-d scatter plot.

The parallel coordinate plot tool contains the following features:

Data filtering – Students define subsets of the data using filters. A filter is simply defined by a collection of minimum and maximum value ranges for each attribute in the dataset. When a filter is first created, the filter minimum and maximum values are set to the minimum and maximum values of the attribute in the dataset. The student is able to adjust these filters, represented visually as small sliders attached to the axes, by simply dragging the pointer up or down along the axis. As a filter slider is dragged, the display is automatically updated – adding or removing observation lines and updating the filter summary in a panel to the right of the plot. For convenience, filter ranges may also be adjusted using a lasso feature which allows the user to drag a rubber band over the area to be included within the filter subset. Examples using the filtering capabilities appear later in the paper.

The tool supports a maximum of eight active filters. For plotting, each filter is assigned a different color. The colors are chosen from the set of colors recommended for nominal groups by Brewer (2003) at Pennsylvania State University. The recommended set of colors can be downloaded in spreadsheet format from: http://www.personal.psu.edu/cab38/ColorBrewer/ColorBrewer_intro.html.

In addition to being able to interactively adjust slider values, as the students explore the data, they may also toggle the hide/show of each filter's observations to temporarily eliminate that filter's observations and allow the student to focus on other subsets of the data. They may also toggle a "Mean/Mode" check box to reduce the plot from all observations to just one line for each filter subset representing the attribute mean or modal values. Figure 4 is a snapshot of the filter control panel.

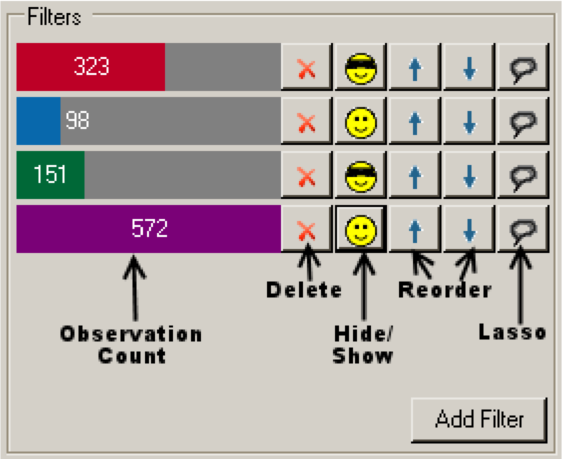


Figure 4: Filter Control Panel

Detail value reading – To facilitate interpretation, given that we had normalized the data values, we displayed the actual minimum and actual maximum values at the top and bottom of each axis. Nominal attributes were represented by equally spaced points listed in alphabetical order from the bottom to the top. To allow more precise value reading of individual points along numeric axes, we implemented a hover feature. A student simply needs to point to a location along an axis and the actual value at that location is displayed adjacent to the pointer. As the student moves up or down the axis, the displayed value changes. As the student moves the pointer between axes, no hover values are displayed. In addition, the user can right-click on an axis location, to pin the value for that point, which will remain visible even after the pointer is moved away from the location.

Representing observation density – To visually represent areas of high observation density, we employed gradient color encoding, similar to the methodology used by Artero *et al*. (2004). Darker shades were used for the low density data points while increasing in whiteness to the higher valued densities. Again the actual color values for shading were chosen from the Color Brewer (2003) recommendations for representing gradients using color. Because in our testing, we didn't feel that the color encoding alone provided sufficient visual resolution of the densities, we added a redundant density encoding by giving the parallel plot a three dimensional representation. The X (left to right) and Y (up and down) display dimensions were used in the same way as traditional two-dimensional parallel plots. We used the Z (back to front) dimension to represent the density of observation values along each axis as an extrusion from the X-Y plane. Figure 14 is an example of a parallel coordinate plot using both gradient color and extrusion to represent density.

Axis reordering - A short-coming of parallel coordinate plots is that in order to evaluate relationships between attributes, those attributes must be visually adjacent. To encourage students to explore different combinations of adjacent attributes, a reordering of the parallel axes was implemented using an interactive dragging of the axes. A key feature of our reordering methodology is that as the column is visually dragged, the plot is continually updated in a smooth animation. There is no single step jump from one ordering to the next that often results in a loss of context to the user and an increase in the cognitive memory load. In the implementation, as an axis is dragged in a horizontal direction to a new location, the adjacent axis that is being approached is also moved, but in the opposite direction, thus allowing the user to visually perceive the swapping of locations of the two columns. To evaluate a given attribute's relationship to each of the other attributes, the student may drag the axis along the full horizontal width of the parallel plot in a single smooth motion, pausing only as each axis is crossed to observe the connecting line pattern.

A VISUAL EXPLORATION OF THE OLIVE OIL DATASET

In this section we demonstrate the use of the above described software in teaching visual exploratory data analysis. The olive oil dataset used contains eight numeric measurements of levels of fatty acid in olive oil samples. There are 572 observations taken from nine different areas in Italy. The nine areas are grouped into three different regions.

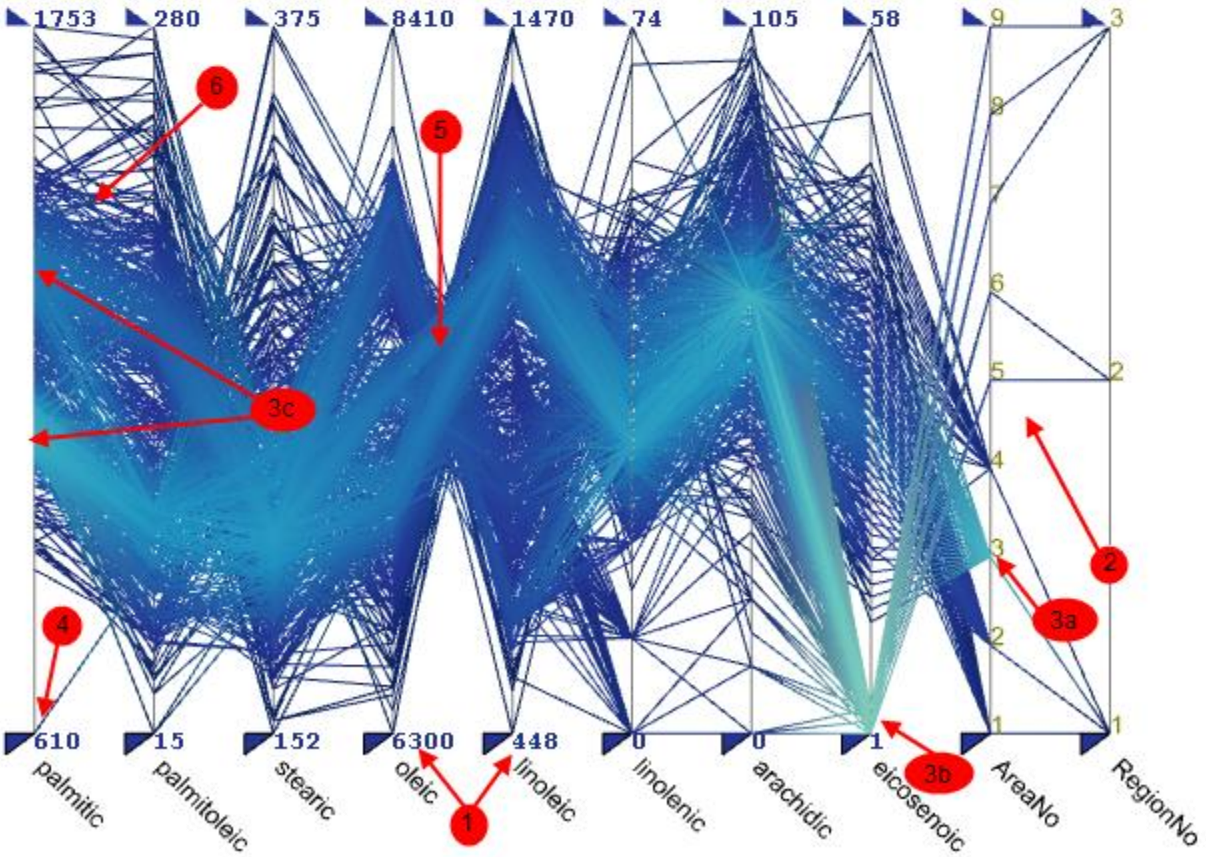


Figure 5: Parallel Coordinate Plot - Olive Oil

We begin the exploration with a parallel coordinate plot of all eight fatty acid attributes plus the region and area. See Figure 5. In this plot, we see a wealth of information:

1. At the bottom and top of each axis are the minimum and maximum values respectively for each acid.

2. There are nine areas numbered 1 through 9 and three regions numbered 1 through 3. By following the connecting lines from area to region, we see that areas 1 through 4 are in region 1, areas 5 and 6 in region 2, and areas 7, 8, and 9 in region 3.
3. By focusing our attention on the color encoding of densities, we see that:
    a. There are more observations in area 3 than any of the other areas,
    b. Most of the observations have a very low level of eicosenoic acid and
    c. The distribution of observations for palmitic acid is bi-modal.

4. There appears to be an outlier with respect to the palmitic acid attribute, given that it has a much lower value than any of the other observations.
5. Oleic and linoleic acid levels appear to be inversely correlated. We see this in the crossing pattern of the lines connecting the two axes.
6. Palmitic and palmitoleic acid levels appear to be directly correlated as the connecting lines are somewhat parallel with the exception of the previously noted outlier and a few other observations visible at the top.

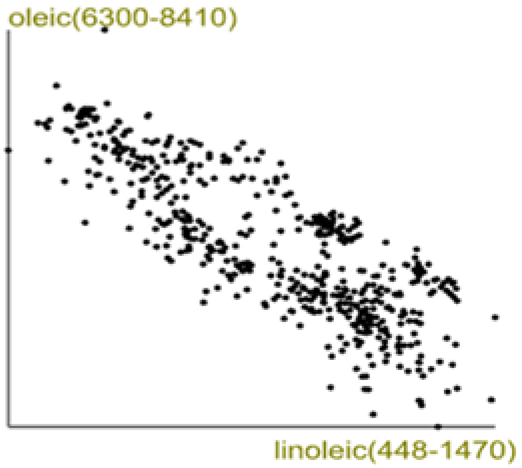


Figure 6: Scatter Plot 2-d

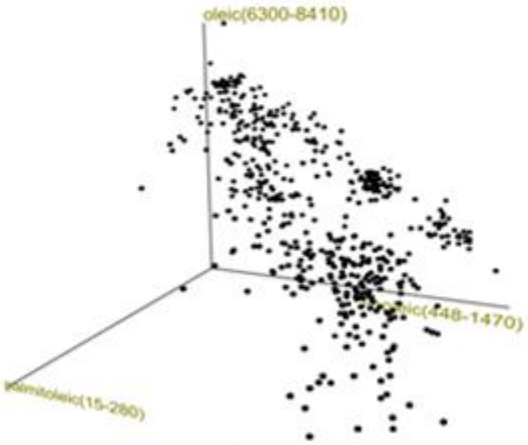


Figure 7: Scatter Plot 3-d

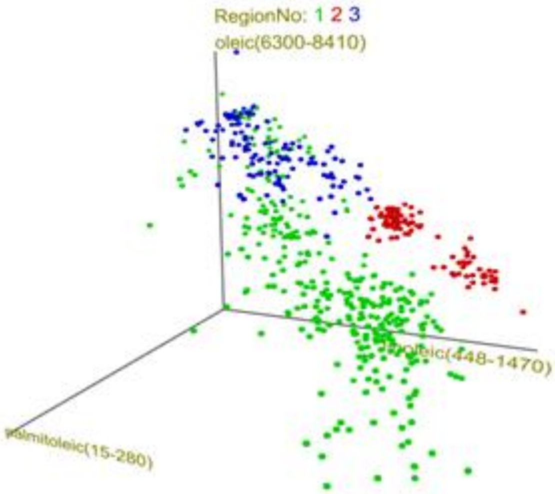


Figure 8: Olive Oil by Region

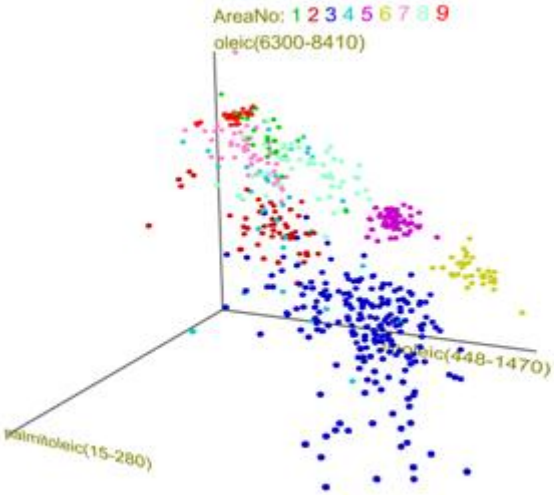


Figure 9: Olive Oil by Area

The relationships noted above (oleic/linoleic and palmitic/palmitoleic) are only visible because of the adjacency of the axes. Students may ask, "Are there other relationships between attributes that we are not currently seeing in the parallel coordinate plot?" To answer this question, they may refer to the color encoded correlation matrix of the same dataset. See Figure 2. The matrix not only confirms the strong inverse correlation between oleic and linoleic acids (manifest by the saturated red shade), but also reveals a strong inverse correlation between oleic acid and both palmitic and palmitoleic acids.

To better understand the nature of the oleic/linoleic relationship, the student can click within the corresponding matrix cell to open a scatter plot of the two attributes. See Figure 6. To expand and explore the tri-variate relationship between oleic, linoleic, and palmitoleic acids, the student uses the select box to add a third axis to the plot for palmitoleic acid. See Figure 7. In viewing this plot, keep in mind that you are seeing a static 2-d projection of a 3-d structure. Perception of the actual structure shape is difficult. Students using the software, however, have the ability to smoothly rotate the image. As they do so, the three dimensional location of the plot points becomes obvious.

As the plot in Figure 7 is rotated, there appear to be two small clusters of observations distinct from the others. To investigate, we color encode our plot using the nominal RegionNo attribute and immediately see that observations in these clusters belong to region 2 (Figure 8).

By changing the color encoding to the nominal AreaNo attribute, students additionally see that these two clusters represent observations drawn from two different areas: 5 and 6 (Figure 9).

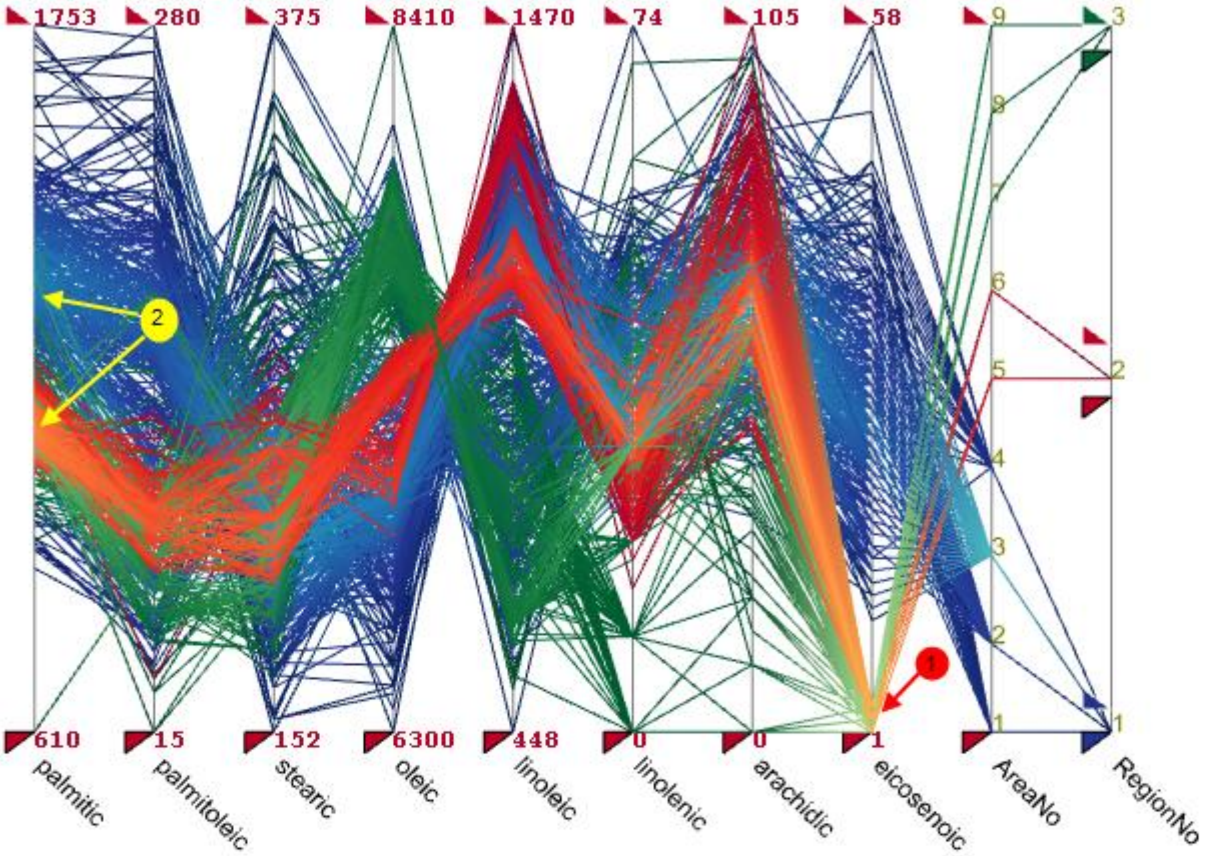


Figure 10: Olive Oil with Region Filters

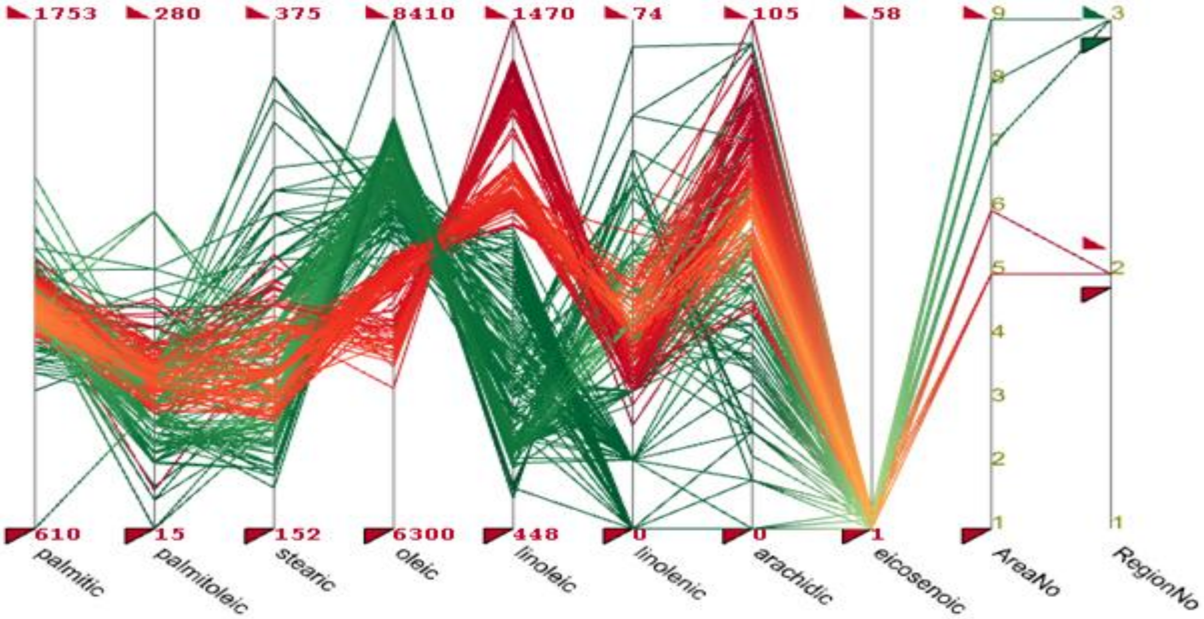


Figure 11: Olive Oil - Regions 2 and 3

Switching back to a parallel coordinate plot, we use the filtering capability to color encode observations for each of the three regions (Figure 10).

Here are represented the differences between acid levels in the three regions. We see:

1. The previously noted concentration of observations at the bottom end of the eicosenoic axis consists of oils from regions 2 and 3 only.

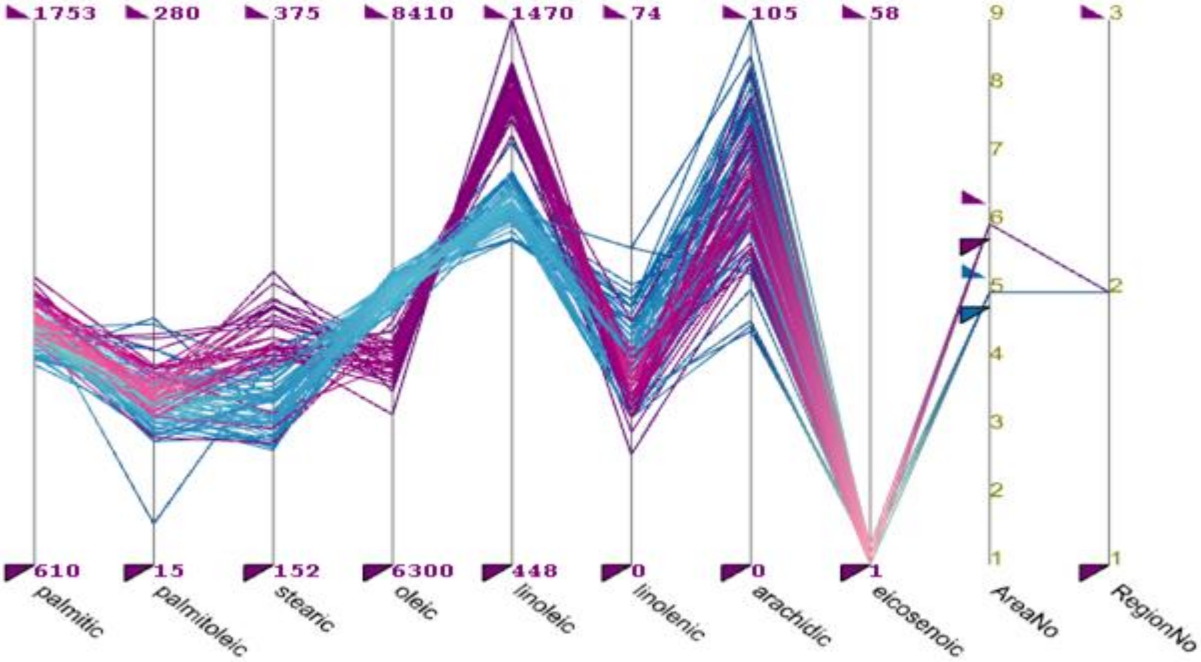


Figure 12: Olive Oil - Areas 5 and 6

2. The bi-modal distribution on the palmitic axis consists mainly of region 2 and 3 observations around the bottom node and region 1 observations at the top.

To focus on the differences between region 2 and 3 observations, we hide the filter for region 1 observations. See Figure 11. Noting again the clustering on the linoleic axis that we first observed using the scatter plot, we modify the filters to compare just areas 5 and 6 within region 2 (Figure 12). This time however, we see concurrently how observations from areas 5 and 6 compare for all eight acids. Specifically we see discrimination between the two subsets using linoleic, oleic, and stearic. We also locate another outlier on the palmitoleic axis when we isolate the area 5 observations.

In a search for additional outliers, students are taught to consider the possibility of bi-variate outliers. That is, those observations which for any single attribute would not be considered an outlier, but in combination with a second attribute are outlying. To do this, we look for connecting line segments between adjacent axes that are in the marginally outlying area with respect to both attributes. Looking back again at Figure 5, note line segments at both the top and bottom of the connections between the linolenic and arachidic axes that could possibly be considered outliers. To investigate, we look at a scatter plot of the two. See Figure 13.

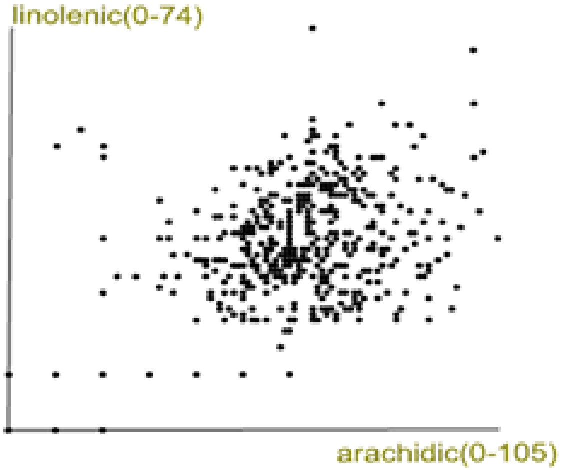


Figure 13: Arachidic vs Linolenic

As expected, we see outlying observations in the top-center and upper-right areas of the plot. In the lower-left area of the plot, we also see a sequence of observations that appear quite suspicious – uniformly spaced points located in an area of the plot where they would likely be considered outliers. We instruct students that when they encounter such distributions in randomly sampled data that they need to investigate the source of these observations.

Due to its relatively small size, the olive oil dataset does not adequately demonstrate the capabilities of the parallel coordinate plot implementation in presenting clusterings within large datasets. To illustrate this capability, we now switch to the out5d dataset, a collection of 16,384 observations along five numeric dimensions. This data was collected in Australia using remote sensing devices (Worcester Polytechnic Institute, 2005).

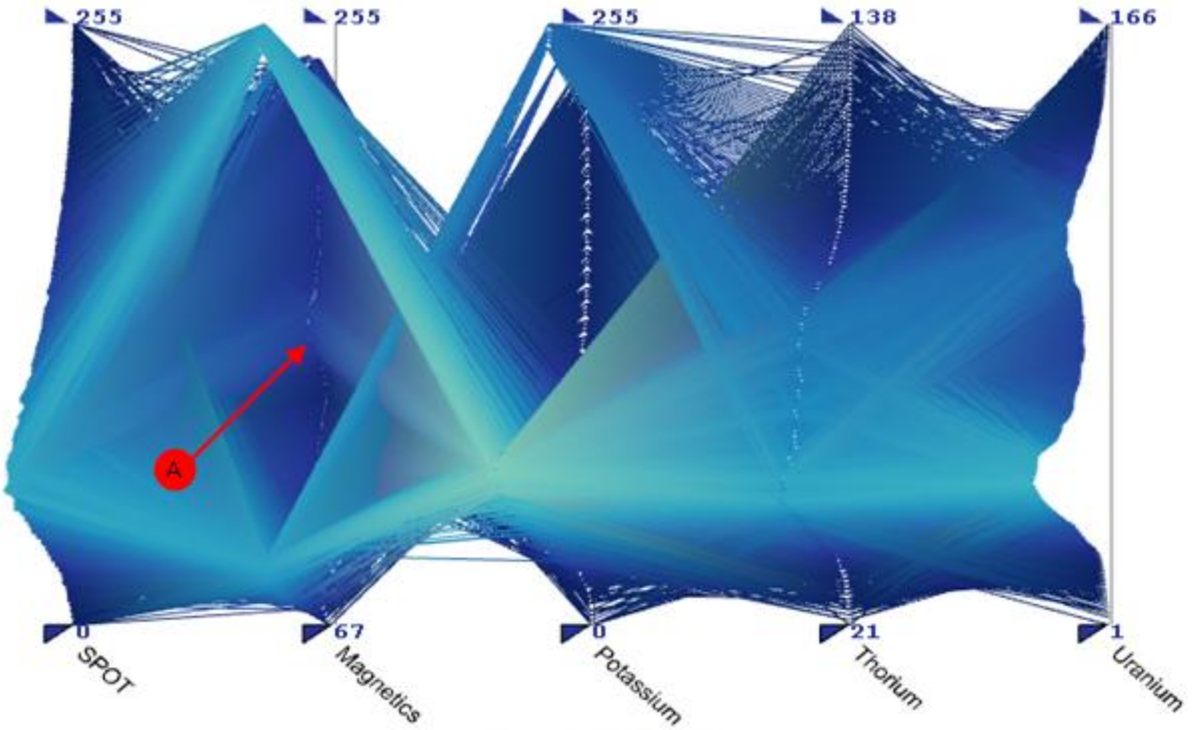


Figure 14: Out5d Dataset

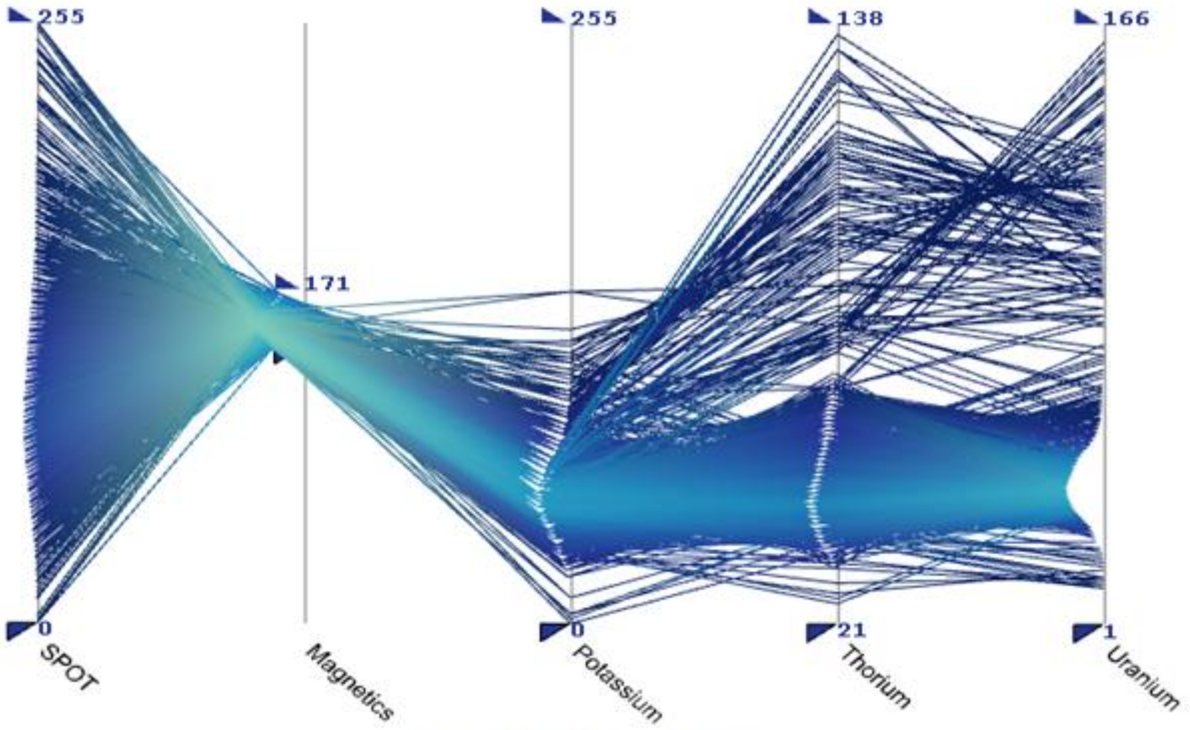


Figure 15: Out5d Dataset Cluster

Figure 14 is a parallel plot of the entire dataset. It has been slightly rotated to reveal the distributions of the SPOT and Uranium attributes. A quick scan of the plot reveals a number of ribbon-like clusters of observations embedded within the entire dataset. On some axes the clusters are visibly distinct (Magnetics for example), while on other axes they merge toward one or two common levels (Potassium for example). To track the clusters through these merge points, we use filtering. For example, suppose that we wish to track the cluster labeled "A" that is visible on the Magnetics axis, but indistinguishable as it passes through the Potassium axis. To do this, we restrict using the filter sliders, to selected observations passing through that visible cluster on the Magnetics axis (Figure 15).

We now see a cluster that is defined independent of the SPOT attribute values, being more precisely defined by the visible ranges of values along the Magnetics, Potassium, Thorium, and Uranium axes. With respect to the scattered observations in the upper ranges of the Thorium and Uranium attributes, students would be encouraged to perform more in-depth analyses in order to identify characteristics of the sub-population from which they were drawn.

SUMMARY OF VISUAL EXPLORATION CAPABILITIES

We began with a list of questions of a "to be explored" dataset that we ask our students to answer in a preliminary exploratory data analysis. In the previous section, we demonstrated the ability of our visual analysis system based on two key plots – parallel coordinate and scatter – to quickly allow students to answer some of those questions. We conclude with a summary of those capabilities.

1c. What are the distributions of individual attribute values?

To assess distributions, scatter plots are effective in datasets small enough that there is not a lot of overlap of the observation values. For large datasets in which the visual points become blob like, observation densities are computed and redundantly encoded using color and height on the Z display axis of the parallel coordinate plots.

2a. Are there outliers in the data?

Uni-variate outliers are readily perceived visually by their separation from the nearest neighbors. Bi-variate outliers are readily perceived on 2-D scatter plots again by their separation toward the corners of the plots. The same is true for tri-variate outliers on a 3-D scatter plot given, that the student has the ability to interactively rotate the display. In parallel coordinate plots, bi-variate outliers are visible as horizontal or nearly horizontal line segments near the tops and bottoms of the plot provided that the two attributes of the bi-variate separation are plotted on adjacent axes and that they are both in the same direction from the attribute mean values.

3. Are there patterns or relationships between attributes or combinations of attributes?

For two or three continuous attributes, the relationships are visually presented using both scatter and parallel coordinate plots. An additional dimension may be included for nominal attributes using color.

4. How do the observations compare? Do all observations appear to originate from the same underlying population or do there appear to be distinct subsets of observations (clusters) within the dataset?

As with previous questions, visual presentation of clusterings are effective using both scatter and parallel coordinate plots for up to three continuous dimensions with a possible fourth dimension of a nominal attribute encoded using color. In our parallel coordinate implementation, because of the density encoding using both gradient color and height, clusterings may be visually perceived beyond three dimensions and with large datasets.

Finally, as we have noted in the summary above, many of the patterns in the data are never visually presented unless the correct attributes are chosen for inclusion in the scatter plots or placed adjacent to each other in the parallel coordinate plots. To help students in their search for these patterns, we found the visual correlation matrix to be a valuable guide in that search.

CONCLUSION

In this paper, we have presented a framework for students in Data Mining courses to conduct an initial exploratory analysis of a dataset. The goal of this analysis is to help students gain an intimate knowledge of the dataset under investigation. We first provide a list of analysis questions for which students are directed to seek answers as they explore the data. We then provide a set of highly interactive visual tools that will allow them to efficiently find answers to these questions.

The interactive features of the tools are important because they encourage a greater level of exploration. Moving from one view to another takes just one or two clicks of the mouse. It quite simply allows students to explore more in the same amount of time compared to using the plot features of a spreadsheet such as Microsoft Excel® or a statistical package such as SAS/Graph®.

The visual capabilities help the students see and assess patterns. Visualizations require less mental effort to assimilate than an examination of tabular or other character based presentations of data. The visual presentations serve to guide students in the next step of data mining – analysis of the dataset using sophisticated statistical and machine learning methodologies. With an intimate knowledge of the dataset, students are able to make proper selection and application of tools, specify how the data is to be used, and correctly set tool parameters. The same visualizations also serve as memory aids in helping students to understand and interpret the results of those analyses.

REFERNCES

1. Amar R, Eagan J, Stasko J, "Low-Level Components of Analytic Activity in Information Visualization," infovis, p. 15, Proceedings of the 2005 IEEE Symposium on Information Visualization (INFOVIS'05), 2005.
2. Artero AO, De Oliveira MCF, Levkowitz H. Uncovering Clusters in Crowded Parallel Coordinates Visualizations. IEEE Symposium on Information Visualization 2004 (Austin, TX, USA), IEEE Computer Society Press; 2004: 81-88.
3. Brewer CA. A Transition in Improving Maps: The ColorBrewer Example. U.S. Report to the International Cartographic Association, special issue of *Cartography and Geographic Information Science* 2003; 30(2): 155-158.
4. Card, S, Mackinlay J, Shneiderman B. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers: San Francisco, 1999; 686 pp.
5. Forina M, Armanino C, Lanteri S. & Tiscornia E. Classification of Olive Oils from Their Fatty Acid Composition. In: Martens H & Russworm H Jr (Eds). *Food Research and Data Analysis*. Applied Science Publishers: London; 1983. 189-214.
6. Inselberg, A. The Plane with Parallel Coordinates, *The Visual Computer* 1985; 1(2): 69-92.
7. Inselberg A. Visualization and data mining of high-dimensional data. *Chemometrics and Intelligent Laboratory Systems* 2002; 60(1-2): 147-159.
8. Moustafa REA, and Wegman EJ. On some Generalizations of Parallel Coordinate Plots. 3/30/2003, http://www.galaxy.gmu.edu/stats/syllabi/IT871/GeneralizedParallelCoordinates.pdf. accessed 4/26/2008.
9. Shneiderman B, Inventing Discovery Tools: Combining Information Visualization with Data Mining. *Information Visualization* 2002; 1(1): 5-12.
10. Tan, P-N., Steinbach, M., and Kumar, V., *Introduction to Data Mining*, Addison-Wesley, 2006.
11. Ware, C. *Information Visualization: Perception for Design*. (2nd Edition) Morgan Kaufmann; 2004; 486 pp.
12. Worcester Polytechnic Institute. 7/15/2005, http://davis.wpi.edu/xmdv/datasets/out5d.html . Accessed 4/26//2008.

NOTES