Visualizing Graduate Admissions Data
CMPS 161: Final Project

Aaron Doubek-Kraft
adoubekk@ucsc.edu

March 20, 2017

Abstract

Abstract goes here!

1 Introduction

In 2017, there were nearly 1100 applicants to the UC Santa Cruz Computer Science graduate program. Given the unusually large size of this applicant pool, making sense of any large-scale trends simply by looking at their records becomes an intractable problem. In addition to the high volume of applicants, the students' records contain a number of potentially interesting variables to be analyzed. This includes quantitative data, including test scores and GPA, and qualitative data, such as the applicants' research interests, countries of origin, and whether or not the student was admitted. In this paper, I attempt to develop visualizations to identify correlations in this data that could be relevant to the selection process. I have attempted to make this paper accessible to a reader with no background in visualization by providing high-level explanations alongside technical descriptions of the problem and my approach.

2 Methods

The large number of variables and the high volume of records to be analyzed makes this a classic multivariate visualization problem, and so I take a relatively standard approach: the parallel coordinate plot. In a typical graph, the number of variables available to be displayed is limited to 2 or 3, the number of spatial dimensions, so only correlations between 2 or 3 variables may be analyzed at a time (for example, in a scatter plot). The parallel coordinate plot generalizes the graphical approach to an arbitrary number of axes, allowing correlations to be analyzed across a relatively large number of variables.

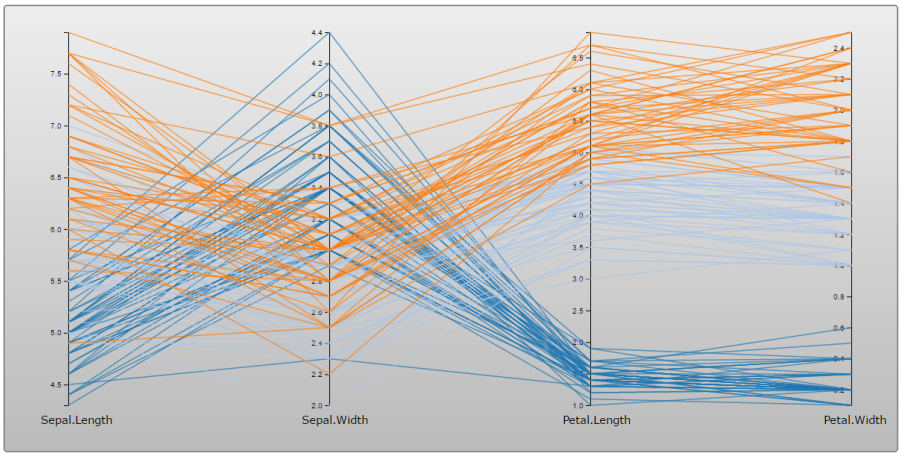

Figure 1: Visualizing Edgar Anderson's Iris dataset. In this example, color is mapped to species of iris.

As an example, I present a standard test case for multivariate visualization:

the iris dataset. Even in the absence of any specific knowledge about irises, some trends are immediately apparent. First, petal length and petal width are clustered for a given species, and they appear to be directly correlated: irises either have large petals or small petals. Similarly, sepal length and sepal width are also clustered, though not as strongly as petal size, and they appear to be inversely correlated (in case you were curious, the sepals are the green parts outside the flower that protect the petals, a fact I had to look up). Now, using the high dimensionality, we can compare axes across the chart, and we can see that sepal length is also correlated to petal width. Now, the strength of this approach becomes clear: it would have taken four or five lower-dimensional charts such as scatterplots to reach the same conclusions that I present here in one condensed figure. I leave analyzing the significance of these trends to the botanists.

3 Implementation

3.1 Processing Dataset

4 Results

5 Conclusion

References

[1] R Datasets [Internet]. Available from: `https://vincentarelbundock.github.io/Rdatasets/datasets.html`

[2] Telea, Alexandru C. Data Visualization Principles and Practice. 2nd Edition. Boca Raton(FL): CRC Press; 2015.

[3] Kosara, Robert. Parallel Coordinates [Internet]. Eagereyes; 2010. Available from `https://eagereyes.org/techniques/parallel-coordinates`

[4] Bostock, Mike. Parallel Coordinates Example [Internet]. d3.js; Available from `http://mbostock.github.io/d3/talk/20111116/iris-parallel.html`