What Does Diversity Look Like?

Tuan Pham Rob Hess Crystal Ju Ronald Metoyer [*]
Oregon State University

Juan Gilbert [†]
Clemson University

Abstract

In this paper we address the problem of visualizing the diversity and depth of a set of objects. We propose a novel visualization for this purpose that is loosely based on parallel coordinates and uses opacity to convey depth information. We evaluate our methods qualitatively using a college admissions dataset and synthetic data.

Index Terms: H.5.0 [Information Systems]: Information Interfaces and Presentation—General;

1 Introduction

In many applications, it is necessary to consider the diversity and depth of a set of objects. For example, in selecting an incoming freshman class, college admissions officials may wish to consider how diverse a particular population of applicants is. In this case, depth information is important as a compliment to information about diversity. For instance, admitting an incoming class of one hundred males and one female would be undesirable, even though both genders are represented in the population.

In most cases, measuring the overall diversity of a set of objects can be decomposed into an examination of diversity in each of a number of separate attributes. For example, in an incoming freshman class, one might wish to explore the diversity of genders, of ethnic backgrounds, of GPAs, etc.

Unfortunately, as the number of attributes and objects to be examined both increase, the number of values that must be considered in gauging diversity increases as the product of the number objects and the number of attributes. This can make assessing the diversity of a large population using only text- or table-based data extremely difficult and possibly enormously time-consuming.

Visually encoding the data offers the potential to overcome this difficulty, but only if a representation is available in which the diversity of the data is readily apparent. For lower-dimensional data, representations such as scatter plots and histograms serve this purpose well, but, as the dimensionality of the data increases, these methods lose their utility quickly. Despite all this, very little work has been done to develop representations that emphasize diversity in a high-dimensional data set, and the methods that have been proposed—such as [8, 3, 4]—exhibit shortcomings that make them unsuitable for visualizing diversity and depth in many important cases.

In this paper, we address this problem using a novel visual representation designed to emphasize both the diversity and depth of a set of objects. To be specific, by diversity, we mean the degree of distribution of objects in attribute space, and, by depth, we mean the degree of concentration of objects in attribute space.

The visual representation we present is loosely based on the classical parallel coordinates representation for multi-dimensional data [2]. Our method differs from traditional parallel coordinates in that objects are represented as series of semi-transparent rectangles, rather than as poly-lines. These rectangles are laid out in such a way that, when objects occupy the same region of attribute space, the corresponding visual region becomes more opaque as the fractional opacity values of overlapping rectangles are added together. Using opacity normalization, we achieve the effect of filling visual space as objects become more uniformly distributed in attribute space. Through this effect, diversity can be assessed pre-attentively.

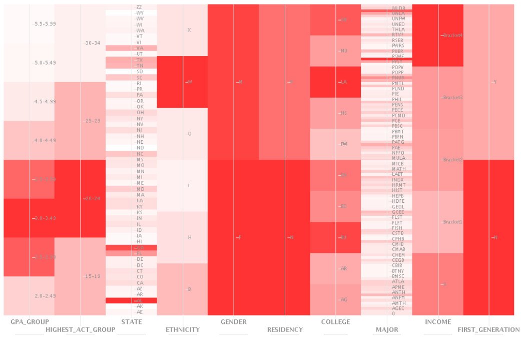


Figure 1: The entire college applicant dataset visualized using semi-transparent rectangles in a parallel axis layout

We apply our methods to the problem of visualizing diversity in a set of college applicants. We use a real dataset containing 2550 applicants (one year worth) to a particular university (Fig. 1). Each applicant is characterized by eleven attributes. We also apply our method to synthetic college admissions data generated to achieve pre-specified values of Shannon entropy, which is used in a variety of domains as a measure of diversity (e.g. [7], [5]).

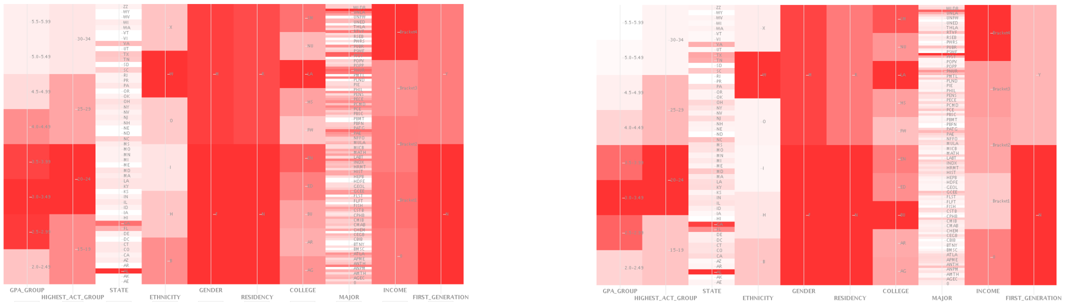
Interestingly, our real dataset was preprocessed using an existing proprietary software package in order to recommend a set of applicants with the specific goal of admitting a diverse incoming class. In addition to gauging our visualization's capacity to answer the questions above, we use the recommendations made by this software package as a baseline against which to compare our method.

2 Design Considerations

Several factors must be considered in designing a visualization to convey diversity and depth information about a set of objects with many attributes. Perhaps the most important of these is choosing effective encodings. Here, diversity and depth are quantitative features of the data. According to the rankings of encoding methods by Mackinlay [6] and Cleveland and McGill [1], position is the most effective encoding of quantitative information. Because diversity, or the distribution of objects in attribute space, is the feature we most want to emphasize, we encode it using position. Specifically, we represent each object as a series of rectangles in 2D space whose positions are calculated using the object's attribute values.

Determining an encoding for depth information is not as straightforward as for diversity. Quantitative encodings, such as length, angle, slope, and area, that are ranked highly in [6] and [1] do not work well for this purpose because they conflict with the position encoding. However, density (or opacity) works as a natural compliment to position to convey the depth of a population of objects. In particular, assigning a fractional opacity value to our object representations allows us to make occlusion between objects work to our advantage, as, in this way, each occlusion serves to increase opacity in the region of attribute space that is common to the occluding

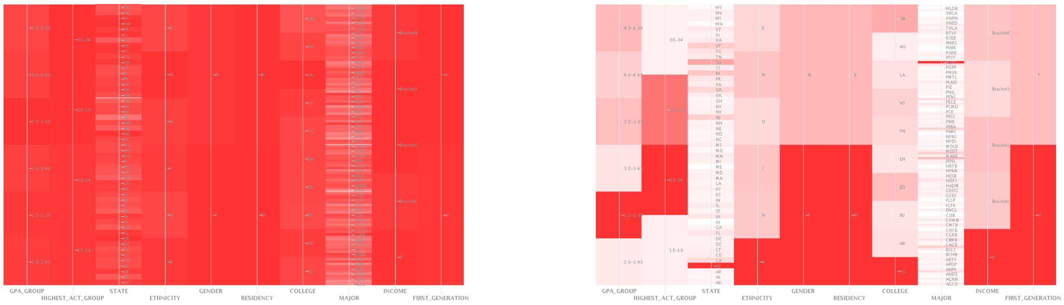
[*] e-mail: [pham|hess|juji|metoyer]@eecs.oregonstate.edu
[†] e-mail: juan@clemson.edu



(a) The subset of recommended applicants

(b) The subset of rejected applicants

Figure 2: The real college applicants dataset, which was analyzed with non-visual software to recommend students based on population diversity



(a) Shannon entropy 20.68 nats

(b) Shannon entropy 10.22 nats

Figure 3: Synthetic college applicant data generated to achieve pre-specified values of Shannon entropy

objects. As a result, deeper regions of attribute space are naturally indicated by visual regions with higher opacity.

3 Visualization

Our visualization (Fig. 1) is a variant of parallel coordinates. Specifically, we adopt the layout of parallel coordinates, in which attribute axes are arranged in parallel. However, in contrast to traditional parallel coordinates, where each object is represented as a poly-line passing through each of the object's attribute values, we represent each object by placing a semi-transparent rectangle on each attribute axis at locations corresponding to the object's attribute values, which are discretized beforehand into buckets. Each rectangle's vertical size is determined by dividing the vertical space on each axis evenly among all of the buckets of the corresponding attribute. Rectangles' horizontal sizes are determined by dividing the allotted horizontal space evenly among all of the attribute axes.

All rectangles for a particular attribute contribute an equal, constant amount of opacity, in RGBA color space, to their corresponding visual locations. Because the range of the alpha channel is limited, we normalize opacity values on a per-attribute basis by assigning an alpha value of zero (full transparency) to buckets not represented by any object and an alpha value of one (full opacity) to the bucket (or buckets) with the highest number of objects within the given attribute. The alpha value for every other bucket is interpolated according to the square root of the number of objects it contains. We have empirically found that interpolating based on the square root of the number of objects results in more recognizable rectangles than does interpolation on a linear scale.

Thus, as objects occupy the same attribute location, the opacity of the corresponding visual location builds to indicate increased depth. The diversity of each attribute can be determined by the distribution of opacity across the buckets on the corresponding axis. Perfect diversity will lead to a space-filling effect, as the buckets on each axis will be colored uniformly. As the population becomes less diverse, some buckets will approach full opacity, while others approach full transparency, resulting in a patchwork effect.

4 Evaluation and Discussion

Fig. 2(a) depicts the subset of college applicants recommended based on diversity by a non-visual proprietary software package, and Fig. 2(b) depicts the subset of rejected applicants. Comparing these two figures and Fig. 1, we can see that the recommended students yield a visualization with a more uniform distribution of opacity, especially in attributes like GPA, major, ethnicity, and residency. This suggests that the recommended applicants are a more diverse population than the set of all students and the set of rejected students. The visual disparity between high- and low-diversity sets is more striking in Figs. 3(a) and (b), which depict synthetic data with Shannon entropy values of 20.68 and 10.22 nats, respectively.

These limited qualitative results are encouraging, though they are far from complete. In future work, we plan to subject our method to a rigorous formal evaluation in which we compare our method to existing methods, such as [8] through user studies. We also intend to explore the effects of parameters such as size, attribute ordering, etc. through additional user studies. Finally, we intend to develop this visualization into an interactive tool with which users may explore and build diverse populations of objects.

References

[1] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387), 1984.

[2] A. Inselberg and B. Dimsdale. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *Proc. IEEE Vis*, 1990.

[3] E. Kandogan. Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In *Proc. KDD*, 2001.

[4] R. Kosara, F. Bendix, and H. Hauser. Parallel sets: Interactive exploration and visual analysis of categorical data. *IEEE TVCG*, 12(4), 2006.

[5] C. Krebs. *Ecological Methodology*. Benjamin Cummings, 1998.

[6] J. Mackinlay. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics*, 5(2), 1986.

[7] L. Masisi, V. Nelwamondo, and T. Marwala. The use of entropy to measure structural diversity. In *Proc. IEEE ICCC*, 2008.

[8] J. Pearlman, P. Rheingans, and M. des Jardins. Visualizing diversity and depth over a set of objects. *IEEE CG&A*, 27(5), 2007.