Using Entropy in Enhancing Visualization of High Dimensional Categorical Data

Jamal Alsakran*  Ye Zhao  Xiaoke Huang  Alex Midget  Jing Yang
Kent State University  Kent State University  Kent State University  UNCC-Charlotte  UNCC-Charlotte

ABSTRACT

The discrete nature of categorical data often confounds the direct application of existing multidimensional visualization techniques. To harness such discrete nature, we propose to utilize entropy related measures to enhance the visualization of categorical data. The entropy information is employed to guide the analysis, ordering, and filtering in visualizations of Scatter Plot Matrix and a variation of Parallel Sets.

1 INTRODUCTION

Existing multidimensional visualization techniques are often underdetermined when directly applied to high dimensional categorical datasets. Such datasets may contain a large number of categorical variables (i.e., dimensions) whose values comprise a set of discrete categories. The discrete nature of categorical data further aggravates the pain coming from cluttering, complexity, and intractability in visualizing large-scale high dimensional data. For instance, bivariate variables (e.g., gender) makes it hard to identify patterns in Parallel Coordinates. On the other hand, traditional visual displays of categorical values, such as Sieve Diagram [5], usually involve only a few variables.

Visualizing categorical datasets has been tackled through different approaches. Sieve Diagram [5], Mosaic Display [2], and Contingency Wheel [1] employ contingency tables in which categories are represented by tiles whose area is proportional to frequency. Parallel Sets [3] improves Parallel Coordinates for categorical data by substituting polylines by frequency based ribbons across dimensions. We propose to use entropy related measures to enhance the knowledge discovery in multivariate visualization techniques, such as Scatter Plot Matrix and a variation of Parallel Sets.

2 ENTROPY RELATED MEASURES

Entropy quantifies the amount of information contained in a discrete data space.

Entropy is computed as:

$$H(X) = \sum_{x \in X} p(x) \log p(x), \quad (1)$$

which provides a measure of the variation, or diversity, of X. It defines the uncertainty of the data dimension. Such information can be harnessed to produce better visual layout.

Joint Entropy is defined over two variables X and Y as:

$$H(X, Y) = \sum_{x \in X, y \in Y} p(x, y) \log p(x, y), \quad (2)$$

where \(p(x, y)\) is the probability of these values occurring together.

Mutual Information measures the reduction of uncertainty of one dimension due to the knowledge of another dimension. It defines a quantity that quantifies the mutual dependence of two random variables. The mutual information is defined as:

$$I(X; Y) = \sum_{x \in X, y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}, \quad (3)$$

Mutual information and joint entropy can lead to better dimension placement and management in visual layout of categorical data.

3 ENHANCING CATEGORICAL VISUALIZATION

3.1 Scatter Plot Matrix

Typically, high dimensional datasets have a large number of dimension pairs, which easily hinder the willingness of data analyzers to evaluate the dataset since they have to browse all Scatter Plots with vague guidance. We visualize the joint entropy and mutual information to guide the discovery of variable relations.

Joint Entropy Matrix: We visualize the joint entropy with colors from low (blue) to high (red), as well as the computed quantities, which is shown in Figure 1(a). The joint entropy is high when the data records distribute more diversely in the corresponding Scatter Plot, and becomes low when the records have lots of overlaps. In Figure 1(a-1), we show the Scatter Plot linked to the highest joint entropy among all dimensional pairs. The particular pair refers to the two dimensions “cap-color” and “gill-color” and the plot shows diverse dot distribution. Figure 1(a-2) displays the plot between the dimensions “gill-attachment” and “veil-type” who has the lowest joint entropy. The joint entropy matrix provides hints for users to conduct their analysis on data variable pairs.

Mutual Information Matrix: In Figure 1(b), the first row is the mutual information between mushroom “class” (edible or poisonous) and all other dimensions. Figure 1(b-1) shows the Scatter Plot with the highest mutual information. It is between “class” and “odor”. The analyzers then can easily find that poisonous mushrooms mostly have the odors like creosote, foul, pungent, spicy and fishy, and on the contrary, most edible mushrooms are no odor and some are almond or anise. Figure 1(b-2) is the Scatter Plot with the lowest mutual information. It shows that mushroom class has no obvious relation with veil-type, because in this dataset, only one veil-type is given.

3.2 Parallel Sets

Dimension management deals with spacing and ordering dimensions in order to produce the best visual layout [6]. We use the entropy information on a variation of Parallel Sets [3] to show how it can help users to manage spacing and ordering of coordinates. The entropy values are also shown in a curve in Figure 2(a). Figure 2(b) is the Parallel Sets visualization of the mushrooms dataset (obtained from the UC Irvine Machine Learning Repository), it includes 8,124 records and 23 categorical dimensions. The category indexing letters are shown in the axes. The colors are defined by the leftmost dimension “class”, where green refers to edible and blue is poisonous.

Sorting categories of neighboring coordinates: We utilize the joint probability distribution, \(p(x,y)\), to sort dimensions categories.

*e-mail: jalsakra@cs.kent.edu
Figure 1: Using joint entropy and mutual information in enhancing scatter plot matrix.

Figure 2: Parallel Sets and entropy chart for Mushroom dataset.

Figure 3: Arranging space between neighboring coordinates and sorting categories with joint entropy.

In practice, between a pair of coordinates, we assume for the left coordinates $d_1$ that its categories are drawn in a given initial order. We choose the order of categories on the second coordinates $d_2$ by using a heuristic greedy algorithm inspired by the work [4]. A category $Y$ of $d_2$ is assigned to an optimal position with regard to the values of all pairs $p(x,Y)$ for any $x$ of $d_1$. Figure 3 shows the result after sorting the dimensions categories.

Arranging space between neighboring coordinates: Assigning more space to coordinates with complex line interactions can help users better perceive details between them. High joint entropy means the diversity of data records between the coordinate pair is high, which usually has complex layout of lines. Thus without changing coordinate order, we compute the joint entropies of all neighboring pairs and use them to arrange the horizontal spacing. Figure 3 shows the result after adjusting the spacing.

Ordering with mutual information: Knowing the left dimension $d_1$, we choose the right dimension $d_2$ as the one having the largest mutual information, $I(d_1; d_2)$, with $d_1$. This pair likely represents more insights than two randomly placed dimensions. Based on the observation, we apply a heuristic greedy method for reordering the coordinates. Starting from the leftmost dimension, we always find the dimension with the largest mutual information, which has not been selected before, as the right dimension. Repeating this process from left to right a re-arrangement of the coordinates is achieved. Figure 4(a) shows the visualization with the original order. Here we color the lines by the left dimension for each section between a pair. Figure 4(b) depicts the result after reordering by mutual information. Here users can find clearer mutual dependency information.

Figure 4: Visualization of the mushroom data with ordering by mutual information. Colored by the left dimension for each neighboring pair.

ACKNOWLEDGEMENTS
This work is in part supported by NSF grant IIS-0915528, IIS-0916131, IIS-0946400, NSFDACS10P1309 and Ohio OBR.

REFERENCES