MULTISET: An Interactive Visual Analysis Tool for Multidimensional Quantitative Data

Liqun Liu and Romain Vuillemot *

ABSTRACT

We introduce MULTISET, a novel visualization technique for analyzing the distribution of multi-dimensional quantitative data, which combines the cumulative distribution function for categorizing quantitative data into categories and a matrix layout for aggregating the elements with the same distributions among generated categories to represent how these elements' multi-dimensional values distribute. Moreover, the attributes of elements in certain group are represented with the following view of matrix layout. In the future, we will also conduct a user study to evaluate the performance of MULTISET.

1 CONTEXT

In many domains, visual analysis of multidimensional data is important to help domain experts understanding data and exploring nontrivial patterns from complex relations. The observation for change of traffic density over time is a key factor to understand the situation of road segments. Therefore, analysis of traffic densities over time is meaningful. For example, **How does the traffic flow of a road segment changes over a day?**

Element ID	Quantitative Variables (Traffic density)				Categorical attributes	Numerical attributes
Road ID	**Time0**	**Time1**	...	**Time23**	**Road Type**	**length**
Road_1	54	43	...	78	A	200
Road_2	78	32	...	84	B	320
Road_3	90	45	...	98	C	580

Figure 1: The structure of road segments data.

Let's consider the scenario for evaluating the traffic status of roads. A road might be fluent at a certain time (e. g., morning) but the traffic congestion happens in the evening as shown in Fig 1. If the traffic flow of roads is diverse widely from time to time, we can not describe the roads with simple words such as smoothly flowing, heavy or traffic jams.

In this situation, the evaluation for the road situation has clutter since a road might be smoothly flowing at 4:00 AM but heavy at 8:00 AM. In order to understand how roads' traffic density change over time and explore the nontrivial patterns from complex situations of traffic densities changing over time, we have to observe both the traffic flow at both peak time and normal time.

Statistical visualizations such as histograms [4] or box plots [3], enable analysts to better grasps the frequency distribution of dimensions of a dataset and proceed with modeling and hypothesis tasks. Statistical charts can even be combined together so that all dimensions frequencies are visually displayed at once. Interactions may be added to further support data selection and aggregation. There has been a significant effort to enhance them to support uncertainty and multi-dimensional data analysis. Moreover, *Mosaic Display* [1] represents the frequency distribution with the vertical and horizontal longs of rectangles. However, those visual techniques have limitations in the scalability of the number of elements and attributes of elements.

*Liqun Liu was with Ecole Centrale de Lyon e-mail: liqun.liu@ec-lyon.fr.

†Romain Vuillemot was with Ecole Centrale de Lyon e-mail: romain.vuillemot@ec-lyon.fr.

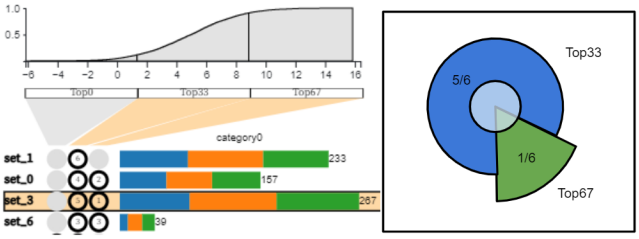

Figure 2: MULTISET is on the left and an illustration of fuzzy sets is on the right. The elements in selected set belonging to Top33 and Top67 simultaneously. There are 5 attributes being scaled as *Top33*, 1 attribute being scaled as *Top67*.

2 MULTISET TECHNIQUE

This method is inspired by an UpSet matrix [2], but the novelty is the set-creation part which is inspired by an interactive cumulative distribution function and MULTISET improved the matrix visualization, which not only represents intersections but combines circles and numbers to represent how many dimensions that an element has attributes in each category. We design MULTISET for visual analysis of distribution for multi-dimensional data. MULTISET is consists of three separated but linked views: 1) cumulative distribution function view, 2) combination matrix. It is illustrated in Fig. **??** ① and ②.

The tool we introduce works as follows. First, the continuous multi-quantitative variables are separated into different categories. And then, based on the generated categories, all the elements are aggregated into the same set if they have same distribution, e. g., road segments would be aggregated if all of them have 3 traffic densities with *Low* category and 2 traffic densities with *High* category. Following with sets view, the categorical attributes and the numerical attributes are visualized with stacked bar chart and box plots following the combination matrix.

2.1 Concept

In possibility theory, a random variable X is taken from the cumulative distribution function. Besides, there is a value x is calculated with a possibility P by the cumulative distribution function. In this case, it means the random variable X have the P possibility being less than or equal to x, the equation is given by:

$$F_X(x) = P(X \leq x) \tag{1}$$

In the cumulative distribution function, it is easy to understand the relative positions of a variable in entire variables. For example, we build a cumulative distribution function for the ages of a people group. If X is a random age, we let is equal to 30 in this example and the possibility is calculated as 0.5. Thus, in this case, we can say there are 50% of people in this group who are less than 30 years old.

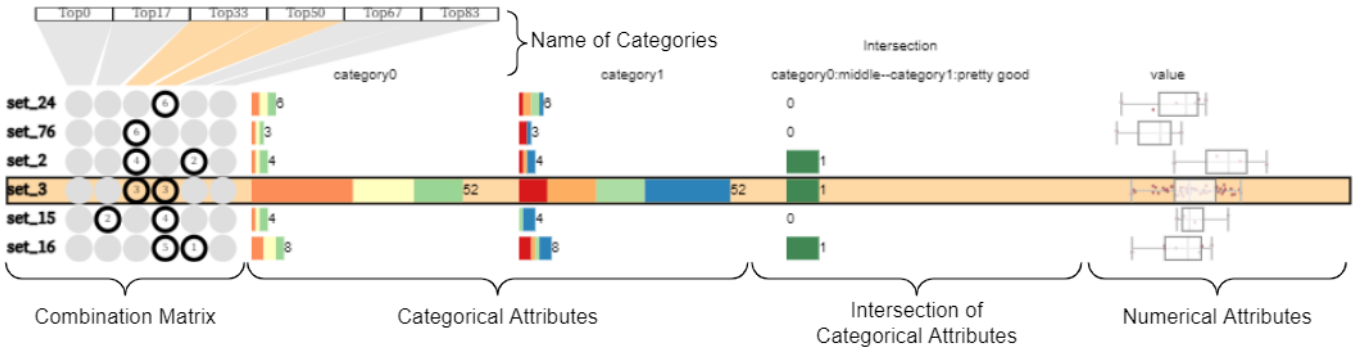

Figure 3: Combination matrix view: categories from cumulative distribution function is visualized as columns in combination matrix, and the rows in combination matrix are aggregated sets that have similar distribution of multi-dimensional quantitative data. The categorical attributes and numerical attributes are represented with stacked bar chart and box plots corresponding to specific sets.

Based on the categories generated from the cumulative distribution function, the values of multi dimensional data should be scaled into different categories and the elements should be aggregated if they have the same distributions in different categories. As illustrated in Fig. 2, ⓝ represent elements in the set have n attributes being scaled in the corresponding category. In contrast, ● means that no attributes of the elements being scaled in the corresponding category. **set_3** is consists of three categories, which means the cumulative distribution function categorize multi-dimensional data into three groups and the count of attributes in each category is represented with the number in circles. It is able to be explained with the figure on the right. The central circle represents all element in **set_3**. The right figure shows there are no atribute belonging to *Top0*, 5 attributes belonging to *Top33* represented with the blue sector, and only 1 attribute belonging to *Top67* represented with the green sector.

2.2 Split Quantitative Values

We design an interactive mapping function based on the theory of cumulative distribution function, as shown in Fig 4 ①. Through this view, the multi-dimensional quantitative data are separated into different categories and these categories are named in a specific way that helps users understand the relative positions of the values in the entire values. Similar to the interactive version of the fuzzy membership function, the number of generated categories are able to be changed.

2.3 Set-based View

The set-based view can both address the set-related tasks by the aggregation of elements with similar distribution such as finding membership degrees and members of a specific set. Besides, the set-based view is also able to address the attribute-related tasks by analyzing the categorical attributes and numerical attributes distribution.

Combination Matrix

In the combination matrix view, as illustrated in Fig 3, all these values of multiple attributes are scaled into categories generated from the cumulative distribution function and then we count how many variables an element has and how these values distribute in categories. The account of attributes distributing in a specific category is represented with a combination of circles and a number (ⓝ). If there is no one attribute distributed in a certain category, it is represented with ●.

Attributes of Sets

In this section, the attributes of sets are divided into two aspects, one is the categorical attributes whose values do not have inherent order and another one is the numerical attributes whose values are continuous. The categorical attributes are represented with stacked bars, colors of which show the discrete values of attributes. As illustrated in Fig. 3, there are two categorical attributes corresponding to sets from the combination matrix. These categorical attributes are visualized with a stacked bar chart, apart from this, there is still an intersection view following with categorical attribute view, which represents the frequency of elements that have specific values in both previous categories. For example, in Fig. 3, the intersection view visualized the cardinality with value of **category0** equal to **middle** and value of *category1* equal to **pretty good**. The numerical attributes are represented with a combination of box plots and scatterplots. The box plot shows the general distribution for values of numerical attributes and the scatterplot represents the individual points for the corresponding sets.

3 Conclusion

In this paper, we introduced MultiSet, which aims at extracting and exploring nontrivial patterns from complex relationships of multidimensional quantitative data. In the future, our works will be focused on as follows

- Add more case studies in different domains and real world dataset;
- Improve the prototype with adding more features such ordering method, etc;
- Conduct an user study to collect feedback from experts in different domains;

References

[1] M. Friendly. Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89(425):190–200, 1994.

[2] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister. UpSet: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1983–1992, dec 2014. doi: 10.1109/TVCG.2014.2346248

[3] D. F. Williamson, R. A. Parker, and J. S. Kendrick. The box plot: a simple visual method to interpret data. *Annals of internal medicine*, 110(11):916–921, 1989.

[4] K. Wittenburg, T. Lanning, M. Heinrichs, and M. Stanton. Parallel bargrams for consumer-based information exploration and choice. In *Proceedings of the 14th annual ACM symposium on User interface software and technology*, pp. 51–60, 2001.

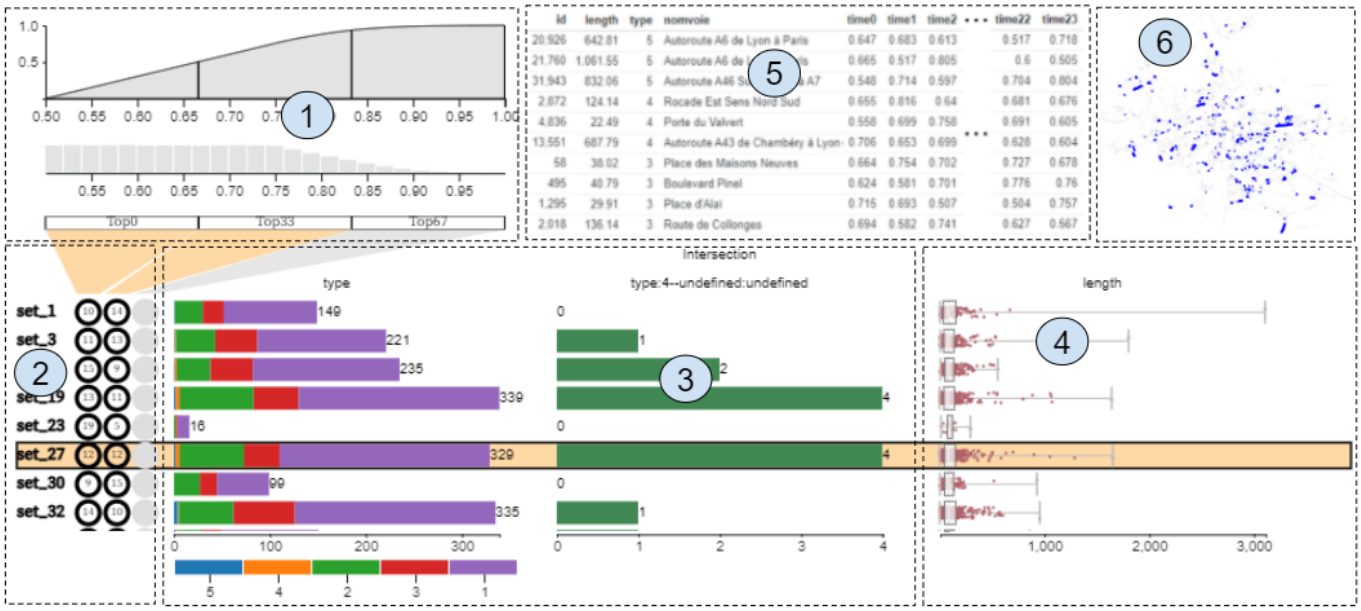

Figure 4: MULTISET first put together all the values of traffic density over time, and then categorize the values into different groups in ① as well as a name is given for each category such as *Top33*. Road segments are gathered based on how many traffic density values being scaled in certain categories in ②. For example, in the row of highlighted **set_7**, the road segments include 24 traffic density values and these road segments have 12 traffic density values scaled into *Top0*, another 12 traffic density values scaled into *Top33* but no traffic density value scaled into *Top67*. The distribution of categorical attributes is illustrated as ③ such as attributed **type** and the specific values of categorical attributes are shown as *Intersection* bar chart. Except for categorical attributes, the numerical attributes of sets are visualized as box charts shown as ④. The elements of highlighted sets (**set_7**) are listed in ⑤ as a table and they are also visualized as a specific plot such as road map in ⑥.