

Visual Mining of Discrete Data

Adalbert F.X. Wilhelm
School of Humanities and Social Sciences
International University Bremen
a.wilhelm@iu-bremen.de

Keywords: Categorical Data, Data Mining, Doubledecker Plots, Interactive Graphics, Mosaicplots.

Abstract

Categorical data sets can be large in three different aspects: they might comprise a huge number of individual cases, they might comprise a large number of variables, or they might comprise a discrete variable with a large number of categories. The term ‘large’ refers in each case to a different level of magnitude: a large number of cases might mean ‘some hundred thousands’ or ‘a million’; a large number of variables might mean ‘more than two’; a large number of categories might probably mean ‘more than five’.

In this talk, we give an overview of the scalability of commonly used visual tools for categorical data, like pie charts, bar charts, and mosaic plots, according to the three aspects of largeness. We discuss the various purposes of visual analysis of discrete data and how the scalability of the display depends on the goal of the analysis. We, in particular, investigate how embedding standard displays in an interactive environment can overcome some of the difficulties with large data sets.

References

- Hartigan, J. and Kleiner, B. (1981). Mosaics for Contingency Tables. In: *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, 268-273, New York: Springer Verlag.
- Friendly, M. (1994). *Mosaic Displays for Multi-Way Contingency Tables*. Journal of the American Statistical Association, 89, 190-200.
- Hofmann, H. (2000). *Different levels of interactivity — Interactive mosaic plots*. *Metrika* **51** (1), p.11-26.
- Hofmann, H., Siebes, A. and Wilhelm, A. (2000). Visualizing Association Rules with Interactive Mosaic Plots. In R. Ramakrishnan, S. Stolfo, R. Bayardo, and I. Parsa (Hrsg.), *KDD-2000 – Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, S. 227-235, ACM, New York.
- Theus, M. and Wilhelm, A. (1998). Counts, proportions, interactions — A view on categorical data. In: *ASA Proceedings - Section on Statistical Graphics*, American Statistical Association, Alexandria, VA, forthcoming.
- Tufte, E. (1983). *The Visual Display of Quantitative Information*. Cheshire: Graphics Press.
- Ward, M., Peng, W., and Wang, X. (2003). Hierarchical Visual Data Mining for Large-Scale Data, Manuscript.