muHVT: Computational Geometry for Visual Analytics

Pravin Venugopal[¶,⋆], **Subir Mansukhani**[¶], **Zubin Dowlaty**[¶]

[¶]Innovation and Development, Mu Sigma Business Solutions Pvt Ltd
[⋆]Contact author: pravin.v@mu-sigma.com

Keywords: Voronoi Tessellations, Vector Quantization, Hierarchical K-Means, Non linear Dimensionality Reduction, Visual Analytics

Given our current capability to store and process vast amounts of data, there is a need to translate data into a visual form. This makes it easy to highlight important features including groups that share common features and outliers along different dimensions of the data.Visual representation of high dimensional data enables users to perceive features in their data quickly thereby augmenting the cognitive reasoning process with perceptual reasoning and enabling the process of generating insight from data to become faster and more directed.To this end, we use techniques and algorithms from *computational geometry* and *nonlinear dimensional reduction* to build a hierarchical "map" of the data and add the ability to overlay features on top of this map in order to visually discover patterns and also validate hypotheses that the user might have about the data at hand.

The muHVT package is a collection of *R* functions for clustering and construction of **Hierarchical Voronoi Tessellations** as a visualization tool to visualize clusters and generate insights from them. The data is compressed in a hierarchical manner to form clusters at various levels using either the *Hierarchical K-means*[1] algorithm where a quantization error governs the number of levels in the hierarchy for a set *k* parameter (the maximum number of clusters at each level) or the *LBG Vector Quantization* (LBG VQ)[3] algorithm which detects the number of clusters for the first level in the hierarchy, based on a specified quantization threshold. The LBG VQ algorithm is useful if the number of clusters in the dataset is not known a priori. The benefit of using hierarchical tessellations lies in the fact that we can analyze data at different levels of granularity. We can think of this as the way digital maps are used, where the user zooms in until the desired information is available. Also, plotting heat maps of the variables in the dataset on the tessellations at various levels of the hierarchy helps in deriving further insight from the data. This next generation segmentation technique gives an edge over the traditional segmentation techniques. For instance, in the example provided in the package we apply this technique to find hierarchical customer segments and overlay the distribution of features across the "map" of the entire dataset in order to understand the characteristics of the different regions of the map. Such datasets are typically found in the CRM systems at Fortune 500 companies. In this package, we use a non linear dimensionality reduction technique called *Sammon's projection*[4] from the **MASS** package. While various distance metrics like Euclidean, Manhattan, Minkowski, etc can be used for computing the distance between the data points, this package also provides a function for the *Jensen-Shannon-Bregman Divergence*[2] distance metric.

References

[1] A Bocker, S Derksen, E. S. G. S. (2004). Hierarchical k-means clustering.

[2] Arindam Banerjee, Srujana Merugu, I. S. D. and J. Ghosh (2005, October). Clustering with bregman divergences. *Journal of Machine Learning Research 6*, 1705–1749.

[3] Linde Y, Buzo A, G. R. (1980, January). An algorithm for vector quantizer design. *IEEE Transactions and Communications 28*, 84–95.

[4] Sammon, J. (1969, May). A nonlinear mapping for data structure analysis. *IEEE Transactions and Communications C-18*, 401–409.