Interactive Poster: Visual Data Mining with the Interactive Dendogram

Peter Imrich[1] Klaus Mueller[1] Raymond Mugno[2] Dan Imre[3] Alla Zelenyuk[3] Wei Zhu[2]

[1] Center for Visual Computing, Computer Science, Stony Brook University
[2] Applied Mathematics and Statistics, Stony Brook University
[3] Environmental Sciences, Brookhaven National Laboratory

Abstract

We describe the interactive dendogram, a visual framework for the mining, survey, and classification of large, high-dimensional data collections. The interactive dendogram is the visual front-end of a classification algorithm providing a hierarchical data decomposition and model refinement capabilities. Hierarchy nodes are placed on concentric circles whose radii are determined by the dissimilarity of the node's sub-tree. A number of features are provided to enable focused viewing, such as collapsible sub-trees, non-linear radial distortion, and fish-eye magnification.

1 Introduction

Classification hierarchies or decision trees are often visualized as static dendograms. We have extended this paradigm into, what we call, interactive dendograms, which offer a variety of user controls to aid in the task of data mining and model development in the scenario of large and time-varying data. Our particular application deals with the survey and analysis of a large collection of millions of digitized aerosol particle spectra. Such a 250-bin mass spectrum is shown in Fig. 1g.

The task of classification of these acquired aerosols based on their spectra is difficult since no comprehensive mass spectrum-indexed library of aerosols exists to date. Given the overwhelming magnitude of the database, we have taken a two-tier approach to enable an interactive approach. To skim the data, we first run an off-line process based on k-means clustering that partitions the data into a few thousand representative groups, each having a classification error of less than the error of the acquisition machine. This procedure calculates the mean spectrum and the covariance matrix for each group and saves them to a file, along with the magnitude, mass and time stamp distribution, and others. The resulting list of "particles" is then utilized in the interactive visual data mining.

The verification of a clustering result will most likely require human expertise, provided by a domain expert, who then either agrees with the clustering or initiates another round of classification with some of the parameter settings changed. Two basic tasks make up this process: (i) evaluation of the present classification results, which is closely related to the task of surveying the information embodied by the data, and (ii) refinement of the model, which basically involves the adjustment of parameters.

{imrich, mueller}@cs.sunysb.edu, {imre, alla}@bnl.gov
{zhu, mugno}@ams.sunysb.edu

2 Related Work

Dendograms are a popular visualization method to illustrate the outcome of decision tree-type clustering in statistics. Usually, these dendograms have a cartesian layout and are drawn as an up-right tree. However, this layout does not make good use of space, it is sparse towards the root and crowded towards the leaf nodes. The spacing between nodes at different levels in the hierarchy is not uniform, which is due to the shrinking number of nodes from bottom to top. For this reason, long, wide-spanning connecting lines are needed to merge nodes at higher levels. A better layout in this respect seems to be the polar or radial layout, where leaf nodes are located on the outer ring and the root is located in the center. A more uniform node spacing results, leading to a better utilization of space and resulting in a better illustration of the class relationships.

3 Our Approach

The radius r of the concentric circle onto which a non-leaf node with dissimilarity *nodeDissim*, obtained from the clustering algorithm [2], is placed is:

$$r = \left(\frac{maxDissim - nodeDissim}{maxDissim - minDissim}\right) \cdot maxRad \qquad (1)$$

where *minDissim* is the partitioning threshold of the k-means pre-processing algorithm, *maxDissim* the dissimilarity of the root node, and *maxRad* the radius of the outer-most circle. We arrange all leaf nodes on the outer-most circle, no matter where the parent node of a sub-tree is located. This is an appropriate placement since each leaf node stands for a partition of the original (large, preprocessed) dataset, and all partitions (or clusters) have been chosen in such a way that their dissimilarities are less or equal *minDissim*.

We have chosen the first layout method of [3], namely the equi-spaced arrangement of leaves on the outer-most circle. We chose this arrangement because it lends itself best for the (future) attachment of node labels. In this configuration, however, when the number of nodes in a sub-tree is large or the tree is highly imbalanced (as is the case in our application), sub-tree edges that connect a node on circle r_1 to a parent node $r_2 < r_1$ may pass through the circle with $r_3 < r_2$, which leads to a cluttered display (Fig. 1a). To avoid this, the angular range of sub-tree edges may not exceed a fan of 180° [1]. By using curved arcs instead of edges this constraint may be can be relaxed (Fig. 1b). However, this may give rise to curved edges that wrap along a large angular distance, destroying the locality of the graph. To reduce

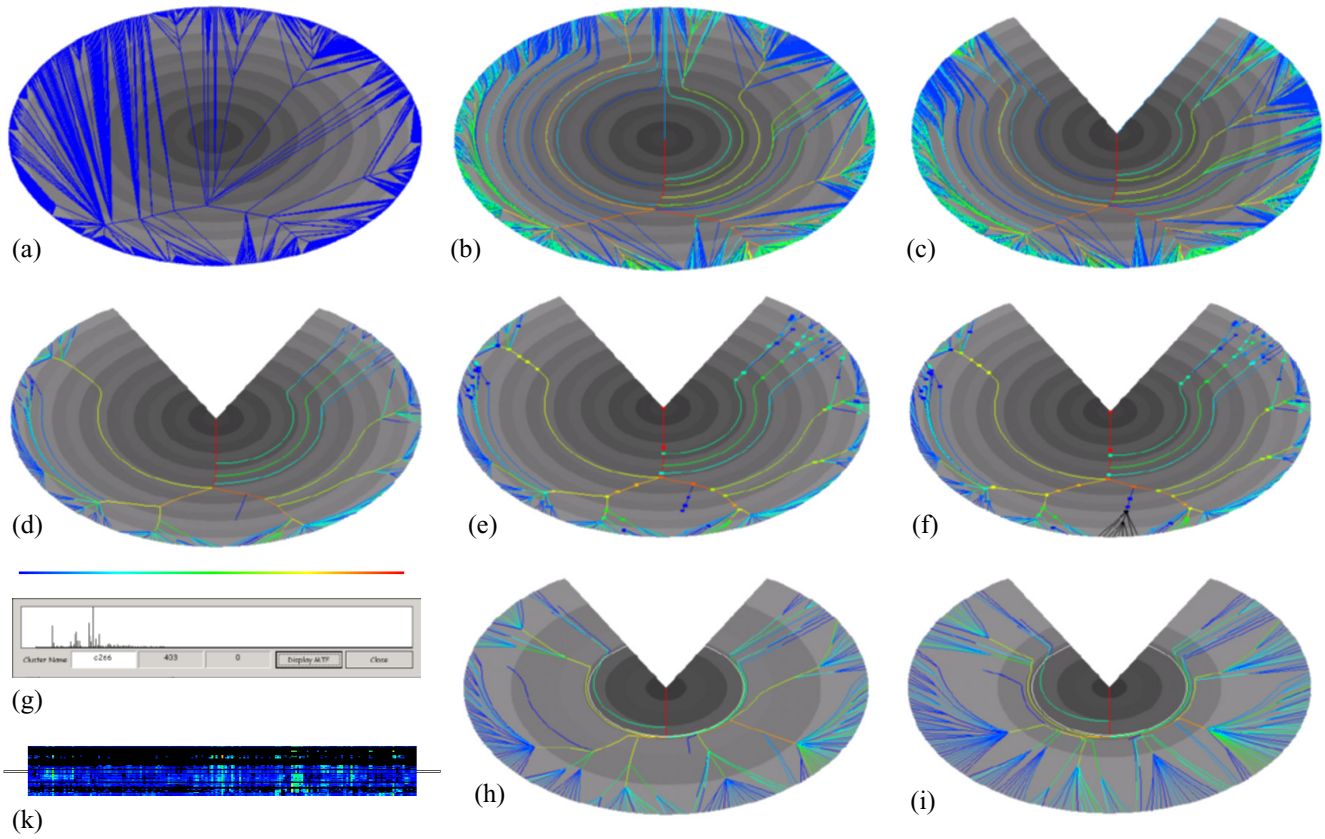

Figure 1: Various features of the interactive dendogram (see Section 3 for a detailed description).

these effects, the user may specify an maximum wrap angle (Fig. 1c).

Edges are colored using a rainbow colormap to indicate the number of particles they carry. In our application, the user may interactively select the minimum number of particles that nodes need to have in order to be drawn. Modifying this threshold automatically redraws the graph with the full rainbow color spectrum mapped to the active node magnitude interval. This pruning of the graph also leads to less curving of the edge arcs (Fig. 1d). The fact that a sub-tree was pruned is indicated by a small knot at the pruned inner node (which is left at the original location in the graph, Fig. 1e). By clicking on the node the user has the ability to expand the node into its full sub-tree (Fig. 1f). A repeat click on the knot will re-collapse the sub-tree.

We also have generalized the mapping of dissimilarity to radius as follows:

$$r = transFunc\left(\frac{maxDissim - nodeDissim}{maxDissim - minDissim}\right) \cdot maxRad \quad (2)$$

where *transFunc* can be any increasing monotonic mapping, established using an interactive curve editor. The mapping can be used to emphasize certain dissimilarity ranges and compress others. We also provide an interactive, piecewise linear stretch function in the dendogram itself where the user first specifies two arbitrary circles in the dendogram. These circles then can be moved apart or brought together which compresses some circular sections in the dendograms and spreads others (Fig. 1h and Fig. 1i). Finally, whenever a node is selected, a window pops up that shows the node's average spectrum, calculated from all nodes in its sub-tree (Fig. 1g).

Aerosol researchers are also very interested in the distribution of particles over time. For this purpose we model the data as a cylindrical stack of dendograms which can be sliced perpendicular to the time axis. The surface of the cylinder is given by the particle distributions in the leaf nodes, i.e., the outer-most circles in the dendogram stack (Fig. 1k).

Using the various viewers in conjunction with the interactive dendogram the scientist can perform the following tasks: (i) label the nodes with aerosol class descriptions, such as "aromatics" or "sulfates", (ii) discard nodes as erroneous measurements, (iii) move nodes along with their subtrees to other parts of the hierarchy, which triggers a refinement of the classification rules.

4 Conclusions

We have described the interactive dendogram, a visual paradigm and application for the mining, survey, and classification of large, high-dimensional data collections.

References

[1] G. Battista, P. Eades, R. Tamassia, and I. Tollis, *Graph Drawing*, Prentice-Hall, 1999.

[2] L. Kaufman and P.J. Roussew, *Finding Groups in Data - An Introduction to Cluster Analysis*, John Wiley & Sons, 1990.

[3] R. Wilson, R. Bergeron, "Dynamic hierarchy specification and visualization," *Information Visualization 1999*, pp. 65-72.