Discovering with Hierarchical Clustering Explorer

Jinwook Seo

Seoul National University

Abstract

Since we publically released Hierarchical Clustering Explorer (HCE) in spring 2002, it has been used for multidimensional data analysis in many domains including biomedical and sociological research. We published a couple of papers on the novel ideas and the technical details about HCE development [1-3]. In this short paper, we will share some stories about the application of HCE to real world problems.

Keywords: multidimensional visualization, HCE.

Index Terms: H.5.2 [User Interfaces]: Screen Design

1 Introduction

In spring 2001, we had a chance to visit a biology laboratory in NIH, where we saw biology researchers looking into a binary tree with genes or skin cancer samples at terminal nodes. The binary tree is called “dendrogram” that represents a hierarchical agglomerative clustering result of either genes or samples. Our meeting with the biologists in NIH (National Institute of Health) allowed us to learn that they were having a hard time deciphering the meaning of the binary tree, using their fingers to figure out clusters embedded in the tree. Hierarchical Clustering Explorer was our solution to the challenge, which later turned out to be prevalent in the field.

The first prototype of HCE was developed in spring 2001 as a term project for Ben Shneiderman’s InfoVis class. A year later, we released the first public version of HCE, which directly addressed the challenge of understanding static dendrograms by implementing dynamic query controls for interactive explorations. In the next version, we added to HCE an interface framework for users to systematically explore 1D or 2D projections of original multidimensional datasets. Several case studies and an email user survey were conducted and we reported the result in a paper for IEEE TVCG [3].

2 Visualization and Interaction Designs

HCE has several visualization components. The main component is the dendrogram view which allows users to interactively explore hierarchical clustering results. Users can drag the minimum similarity bar to interactively separate clusters by cutting the branches of the dendrogram that meets the bar. They can also drag the detail cutoff bar to see the average patterns of branches below the bar.

Another important component is the rank-by-feature framework, which enables users to systematically explore 1D and 2D orthogonal projections of the original multidimensional data using easy-to-understand visualizations such as histograms and scatterplots. These two main components as well as others are coordinated with each other. For example, when users click on a cluster at the dendrogram view, the selected items in the cluster are highlighted in scatterplots and other views.

The visualization and interaction design considerations are presented in more detail in our previous publications [1, 2].

3 Discovery Process and Gained Insights

In addition to the fact that HCE helps biologists to better understand their hierarchical clustering results of genes or samples, it is interesting to note that they can gain better understanding of how the clustering algorithm works. It was surprising to know that a professor at a business school (University of Maryland) used HCE to teach how the hierarchical agglomerative clustering works. As shown in Figure 1 (from top to bottom), students can learn how the algorithm merges smaller clusters to make bigger ones by gradually dragging up the minimum similarity bar.

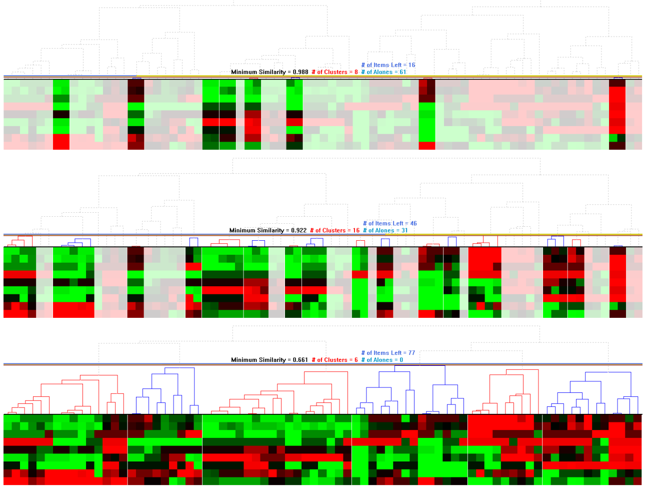

Figure 1. HCE: interactive cluster separation with the minimum similarity bar helps users learn how the hierarchical clustering algorithm works.

Since HCE was developed for microarray data analysis at the beginning, it is well-known to microarray researchers. There are quite a few biology journal papers that cited HCE as an analysis tool for their microarray data. The most prevalent usage pattern in this field is that users play with the minimum similarity bar (and detail cutoff bar for large datasets) until they see a meaningful separation of clusters. Then they identify clusters that deserve further investigations: the ones with focus genes or the ones with target genes selected by interactive search in other views such as the profile search view in HCE. Once they find such clusters, they often generate a hypothesis that genes in the cluster might have similar or related biological functions to the focus/target genes.

Using this discovery process, a team of molecular biologists at the Children’s Research Institute could identify 18 genes involved in the muscle regeneration process as shown in Figure 2. Among other clusters of importance, they mainly focused on the 3 day

jwseo@cse.snu.ac.kr
School of Computer Science and Engineering
Seoul National University
599 Gwanak-ro Gwanak-gu, Seoul 151-744, Korea

cluster where a focus gene (MyoD) belongs. Then they moved on to 12 hour clusters where other focus genes are.

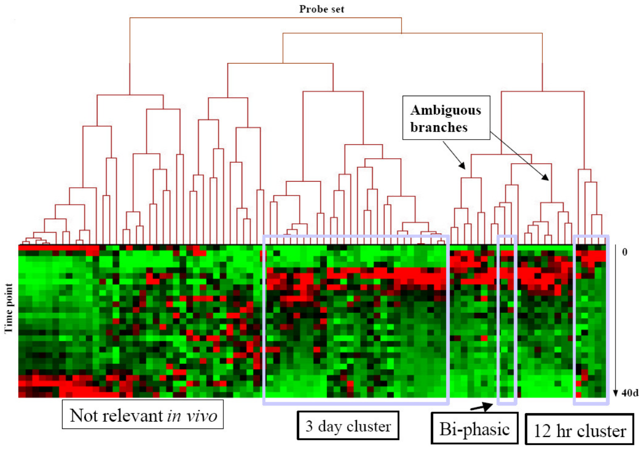

Figure 2. Identifying genes related to the muscle regeneration. 3 day cluster and 12 hour cluster were selected for further investigation.

Not only did microarray researchers use HCE, but researchers from other fields also used HCE for their discovery tasks. An interesting example is the analysis of a large archive of personal email communications. It is interesting to note that multiple-view coordination plays an important role when users try to figure out why a group of items (emails in this case) are clustered together. For example, in the following figure (Figure 3), the pattern clarifies itself in the profile chart view when users click on a cluster in the dendrogram view; the selected cluster represents the relationship that was very active in 1988.

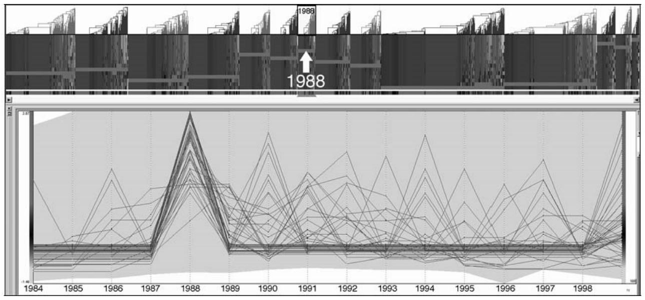

Figure 3. Finding patterns in a personal email archive using coordination between dendrogram view and profile search view.

Through users' comments and case studies, we identified many improvement possibilities in HCE as well as some errors in users' datasets. For example, in a case study, we had to add an option to exclude missing values not to distort the real relationship because missing values were encoded as zeros (Figure 4). A small fix like this one significantly improved the overview of the overall score distribution to clearly reveal distinctive blocks of strong linear relationships.

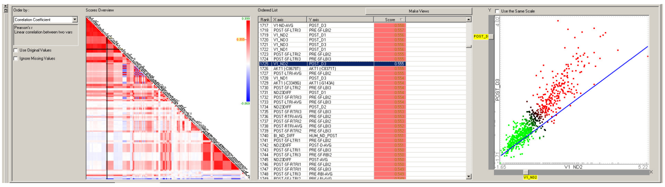

Figure 4. Rank-by-feature framework, revealing erroneous encoding of missing values.

An interesting and unexpected usage pattern was observed when we did a case study with a meteorologist. He wanted to explore his large multidimensional dataset of aerosols to figure out a good way to classify them according to their characteristics. After playing with the minimum similarity bar, he turned to the rank-by-feature framework to find out a couple of interesting quadratic or linear relationships between variables (or dimensions). The unexpected part is that he then started selecting clusters in the dendrogram view to check whether they strengthen or break down the relationship (Figure 5). He actually identified a couple of interesting aerosol clusters that might be useful to improve the underlying model regarding the relationship between aerosol concentration and the amount of water vapor.

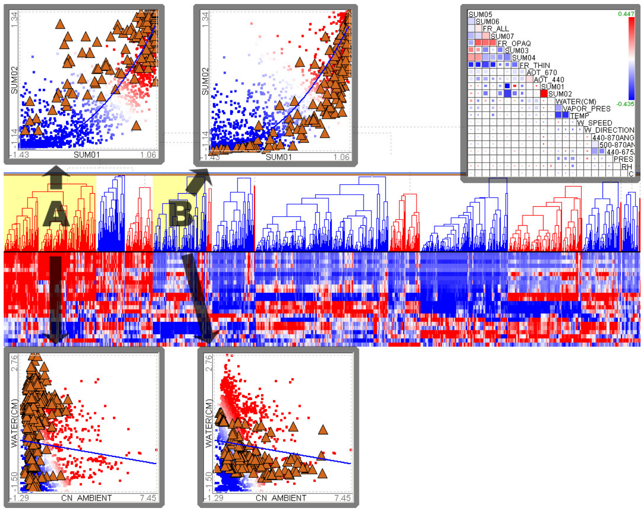

Figure 5. Coupling of the dendrogram view and the rank-by-feature framework.

4 The Impact and Future of HCE

Our email user survey told us that HCE improved the way most users analyze their data sets at least somewhat significantly. For example, a corporate development manager at a company commented: "We performed clustering and - based on the HCE output - modified our specifications for a software product that we offer to non-profits. Very direct link between the HCE usability and good cause!"

There are many microarray researchers who use HCE as their primary tool for microarray data exploration. Even though we could not maintain or improve HCE as well as we did before, we still can see some users download HCE every weekday and some others ask for more functionality.

Fortunately, we have recently got a small funding to develop a newer version of HCE.

References

[1] J. Seo and B. Shneiderman. Interactively Exploring Hierarchical Clustering Results. Computer, 35(7):80-86, 2002.

[2] J. Seo and B. Shneiderman. A Rank-by-Feature Framework for Interactive Exploration of Multidimensional Data. Information Visualization, 4(2);808-814, 2005.

[3] J. Seo and B. Shneiderman. Knowledge Discovery in High Dimensional Data: Case Studies and a User Survey for the Rank-by-Feature Framework, IEEE Transactions on Visualization and Computer Graphics, 12(3):311-322, 2006.