TECHNICAL RESEARCH REPORT

Knowledge Discovery in High Dimensional Data: Case Studies and a User Survey for an Information Visualization Tool

by Jinwook Seo and Ben Shneiderman

TR 2005-100

Knowledge Discovery in High Dimensional Data:

Case Studies and a User Survey for an Information Visualization Tool

Jinwook Seo and Ben Shneiderman
{jinwook, ben}@cs.umd.edu

Human-Computer Interaction Laboratory &
Department of Computer Science
University of Maryland, College Park, MD 20742

Abstract

Knowledge discovery in high dimensional data is a challenging enterprise, but new visual analytic tools appear to offer users remarkable powers if they are ready to learn new concepts and interfaces. Our 3-year effort to develop versions of the Hierarchical Clustering Explorer (HCE) began with building an interactive tool for exploring clustering results. It expanded, based on user needs, to include other potent analytic and visualization tools for multivariate data, especially the *rank-by-feature framework*. Our own successes using HCE provided some testimonial evidence of its utility, but we felt it necessary to get beyond our subjective impressions. This paper presents an evaluation of the Hierarchical Clustering Explorer (HCE) using three case studies and an email user survey (n=57) to focus on skill acquisition with the novel concepts and interface for the rank-by-feature framework. Knowledgeable and motivated users in diverse fields provided multiple perspectives that refined our understanding of strengths and weaknesses. A user survey confirmed the benefits of HCE, but gave less guidance about improvements. Both evaluations suggested improved training methods.

Keywords: Information Visualization Evaluation, Case Study, User Survey, Rank-by-Feature Framework, Hierarchical Clustering Explorer

1 Introduction

The Hierarchical Clustering Explorer (HCE, available at www.cs.umd.edu/hcil/hce) is an interactive knowledge discovery tool for multivariate data, especially of microarray data sets [19]. Its unique visualization interface and powerful analytic tools, based on more than three years of effort, have induced almost 3000 downloads since April 2002. In addition to our genomic research papers with biologist partners and our information visualization publications, we are encouraged that at least six scientific papers from authors unknown to us were published since 2004 that describe using HCE in their analysis [2, 3, 6, 12, 16, 26]. This gives us encouragement that HCE is useful, but we wanted to understand its strengths and weaknesses in a more focused manner. This paper describes the maturation of HCE as guided by user needs, and offers two evaluation strategies, case study reports and an email user survey, to assess the strengths and weaknesses of the rank-by-feature framework as implemented in HCE.

Our early work focused on implementing hierarchical clustering with an interactive interface to support exploration. Based on feedback from initial HCE users, we realized that the clustering results and dendrogram were helpful, but that integration with other representations of multivariate data sets would greatly increase HCE's value. We added 1-D histograms, 2-D scatterplots, parallel coordinates, tabular data, and a Gene Ontology viewer, all as coordinated windows so that selections in one window would produce highlights in all windows.

Since the use of multiple windows could become overwhelming, we interacted with users in many fields, to develop a set of guiding principles. The GRID principles (Graphics, Ranking, and Interaction for Discovery) offer a strategy for analyses of multivariate data sets using low dimensional projections [21]: (1) study 1D, study 2D, then find features, (2) ranking guides insight, statistics confirm. This more structured strategy extended common recommendations in

exploratory data analysis with the goal of replacing opportunistic discovery by a more orderly process that was thorough and repeatable.

GRID principles encourage analysts to clarify their goals first and to apply appropriate computational methods as ranking criteria to rank all possible 1D or 2D projections. In this way, more thorough exploration of multidimensional data sets becomes possible. GRID principles have been implemented in HCE as a user interface framework called the *rank-by-feature framework*, where users can select a ranking criterion, see the graphical color-coded overview of the ranking result, and interactively explore all axis-parallel 1D or 2D projections of a multivariate data set. Since 1D projections are presented by histograms, 1D ranking is called *Histogram Ordering*. Since 2D projections are presented by scatterplots, 2D ranking is called *Scatterplot Ordering*. Similar but separate graphical interfaces were designed for both rankings (Figure 1 & Figure 2). Available ranking criteria include Pearson correlation coefficient, regressions, uniformity, normality, number of outliers, and size of the biggest gap. Detailed explanation on the ranking criteria and the user interface design is presented in [21].

An early version of HCE (version 2.0 without the rank-by-feature framework) was successfully used with our biology collaborators in two projects with gene expression data. We proposed a general method of using HCE to identify the optimal signal-to-noise ratios in Affymetrix gene chip data analyses [17, 18]. HCE's interactive features helped researchers find the optimal combination of three variables (probe set signal algorithms, noise filtering methods, and clustering linkage methods) to maximize the effect of the desired biological variable on data interpretation. HCE was also used to analyze in vivo murine muscle regeneration expression profiling data using Affymetrix U74Av2 (12,488 probe sets) chips measured in 27 time points.

HCE's visual analysis techniques and dynamic query controls played an important role in finding 13 novel downstream targets that are biologically relevant during myoblast differentiation [27].

Our pride in HCE was bolstered by supportive email notes that told us that HCE could handle data sets too large for other software packages, and by enthusiasm for the interactive features. While such testimonials are appreciated, we sought to evaluate HCE by more traditional scientific strategies to produce more generalizable and authoritative results. In an impressive evaluation of four software tools as used by 30 professional biologists, Saraiya *et al*. evaluated HCE with other major microarray visualization tools. HCE outperformed other tools in enabling users to make significant insights with the Viral data set [7], although learning problems lowered HCE's performance in other tasks. Our two projects and Saraiya *et al*.'s evaluation showed the overall usefulness of HCE, but the rank-by-feature framework was not evaluated in these studies.

Evaluating an information visualization or knowledge discovery tool can help identify usability problems and validate an innovative design idea. Therefore, we set out to evaluate the rank-by-feature framework and its implementation in HCE 3.0. Our goals were to understand user difficulties in learning the rank-by-feature framework and the HCE 3.0 interface. We hoped to improve the interface and our training methods for new users. While controlled experiments provide rigorous results, they are not appropriate for situations in which lengthy learning times, extensive domain knowledge, and diverse work styles are expected. Furthermore, our goal is not to prove narrow hypotheses, such as statistically significant differences in performance speed, but *to demonstrate benefits for knowledge discovery in research-level tasks*. This paper describes these new evaluation results using 3 detailed case studies over 8 weeks and an email user survey with emphasis on the rank-by-feature framework implemented in HCE 3.0.

We started five new case studies with researchers in biology, statistics and meteorology. Three case studies were finished with valuable results, but two others were terminated because one researcher changed jobs in the middle of study and the other's expectation from the case study was not compatible with the evaluators. Two case studies were done in the Hoffman Lab at the Children's National Medical Center. One case study was done with a meteorologist at the University of Maryland, College Park.

The objective of these case studies was to show the usefulness of HCE and the rank-by-feature framework in realistic research tasks. The main question that we hoped to answer with was "How do HCE and the rank-by-feature framework change the way researchers explore their data sets?" Participating researchers have primarily used text-based analysis tools or tools that produce static visualizations. Our case studies, summarized in section 3, provide strong support for the usefulness of HCE and the rank-by-feature framework.

Even though intensive case studies with a small number of subjects can show the usefulness of a system and idea, a larger scale user survey may provide more generalizable results. After analyzing the HCE download log and users' comments from email inquiries, we designed a user survey. About one third of the users who have downloaded HCE since April 2002 indicated their intended use of HCE. Using that information, a email questionnaire was sent out to identifiable users. The user survey results are discussed in section 4.

2 Related work

Typical user studies for the evaluation of information visualization tools have been done in tightly controlled laboratory settings where predefined tasks based on a small number of data sets are performed within an hour or two. These evaluation methods are suitable for understanding the potential and limitations of specific features of an information visualization tool. Reviews

and surveys of such empirical evaluations can be found at [4, 5]. Lieberman's arguments against controlled experiments in his CHI 2003 Fringe session, "The Tyranny of Evaluation [11]," emphasize the inherent variability of human subjects and the number of variables to control [9]. For advanced information visualization tools we may also raise concerns about laboratory studies since: (1) researchers rarely start with clearly defined tasks, i.e. part of their work is discovering what questions to ask, (2) researchers must learn to reformulate their data analysis strategies to accommodate new tools, and (3) exploratory data analysis may take place over days or weeks.

Saraiya *et al.* tried to combine the benefit of controlled experiment and usability testing in by quantifying insights –individual observations about the data by the participant [15]. Their method can help microarray analysts choose a right tool, but the short training time (15 mins) for all four tools could introduce some bias since some tools might require much more time to get accustomed to.

The challenge of information visualization evaluation has recently drawn attention from many researchers. A promising outcome is the organization of information visualization contests and the compilation of benchmark data sets and tasks. Other possible steps are to conduct longitudinal case studies and report success stories so that designers can understand problems and potential users can gauge efficacy [13].

Longitudinal case studies are performed with typical users exploring their data sets in their familiar working environment for days or weeks. Case studies also known as "workplace studies" or "field studies" could reveal how information visualization techniques change the way users perform their analysis tasks. For example, Gonzalez *et al*. show in their long term (>6 weeks) workplace study that data analysts can benefit from information visualization systems when the systems are redesigned to be complementary products of current workflow systems [8].

These evaluation methods also have their limitations. Since one situation cannot be duplicated, the experimenter may not get the same results in a different situation. Even though participants may be impressed with the tools being tested, there might be other tools that could be even more beneficial. For the evaluation results to be generalizable to other situations and convincing to potential users, it is necessary to compile more evidence through multiple case studies in multiple fields of research. Even though there is no evaluation that will guarantee success for the other users with differing needs, multiple case studies and testimonials can inspire confidence and increase understanding of what features are especially effective.

To address some of these concerns, we conducted longitudinal case studies with users from different fields including a biologist, statistician, and meteorologist. To reach out to a more diverse user population, we also conducted an email user survey on the usage of HCE and especially on the usefulness of the rank-by-feature framework.

3 Case Studies

One of the research labs that most intensively use HCE is the Hoffman Lab at the Children's National Medical Center in Washington, DC. We have been members of the bioinformatics team there and attended the biweekly team meeting for two years. The first author's major role in the lab was as a consultant who helped researchers analyze their data sets with HCE and sometimes other tools. Researchers in the lab have been using HCE for Affymetrix GeneChip analysis since the summer of 2002. We successfully finished two case studies in the lab with a biologist and a statistician. To make our study more general and authoritative, among many other HCE users in non-biology fields, we recruited one motivated user in the meteorology department at the University of Maryland, College Park. All participants: (1) are motivated, (2) had not used any interactive data exploration tool like HCE before, (3) have their own favorite tools for the

research and analysis, which are mostly text-based and not interactive, and (4) are at the early stage of data analysis. In this section we report the results from case studies with these three participants.

3.1 Methods and Goals

The main methods of these case studies were participatory observations and interviews. While observing and interviewing these researchers, we also helped them learn to use HCE and when necessary improved HCE according to their requirements. It was a rapid interactive iteration process where important requests were implemented during the study period and then observations and interviews were conducted again using the improved system.

For each participant, we arranged a weekly meeting for 4-6 weeks. Although sessions were originally scheduled for thirty minutes, they usually lasted more than an hour because of prolonged discussion of problems and findings during the session. At the first meeting, we intensively taught participants how to use HCE with many examples including small general data sets and large data sets of specific interest to the research. After each meeting, participants were asked to use HCE in their everyday work. Between sessions we communicated via email or phone conversations. During the session, we sat by a participant and observed the participant using HCE, collected their implementation requests, and asked a series of questions to better understand their findings and to examine their experience with HCE. At the end of each case study, the researchers wrote a short final report on their experiences with HCE. Interestingly two of them voluntarily sent us their report without any request. In the report, they usually included screenshots to illustrate interesting findings, and noted comments on the findings.

These three case studies were focused on the evaluation of usefulness of HCE's tools, especially the rank-by-feature framework. The observations and interviews were focused on the

following aspects: (1) how does HCE improve the way users analyze multidimensional data sets, (2) how does the score overview help users identify interesting projections, (3) how does the histogram/scatterplot browser help users traverse projections, (4) what are the most frequently used ranking criteria, and (5) Identify possible improvements in HCE and the rank-by-feature framework.

The next three sections describe case studies with the molecular biologist (P1), statistician (P2), and meteorologist (P3), respectively.

3.2 Affymetrix Data Set with Three Cell Types

A molecular biologist (P1) used one of the accepted animal models for acute lung injury to study inflammatory and immunological events occurring as a result of an LPS (lipopolysaccharide) injection which induces a systemic infection in a model system. P1 performed an Affymetrix microarray project with 12 samples, 4 samples for each of 3 cell types (TH1, TH2, and Platelet) from mice. TH stands for T-helper cell (immune cells). TH1 cells are active in cellular immunity and TH2 cells are active in humoral immunity. Both mature from a common precursor TH cell. The balance of each type of TH cell present in the body seems to be important in determining the progression and outcome of various disease states. Mice were injected with LPS and sacrificed after 0, 24, and 48 hours. P1 monitored the gene expression of these peripheral blood cells.

Through an interactive optimization of signal-to-noise ratios in HCE [17], P1 decided to use the MBEI algorithm available in the dChip application [10] to calculate gene expression values from the Affymetrix CEL files. Most tasks P1 performed with HCE was exploratory. P1 wanted to build meaningful hypotheses and to find a small number of genes that worth further investigation.

3.2.1 Histogram Ordering

As most users do with HCE, P1 also tried the histogram ordering first after loading the data set and looking at the dendrogram view. Among available ranking criteria, the "biggest gap" ranking held the most immediate interest for her. P1 was intrigued by the fact that gaps reveal interesting outliers. Figure 1 shows a ranking result by the size of the biggest gap. The selected histogram clearly shows an outlying probe set in the sample (48_1_TH2), which was identified as having the second largest gap. This probe set was similar to "A kinase (PRKA) anchor protein (yotiao) 9" which is a cytoplasmic/centriolar protein having protein-binding and kinase activity.

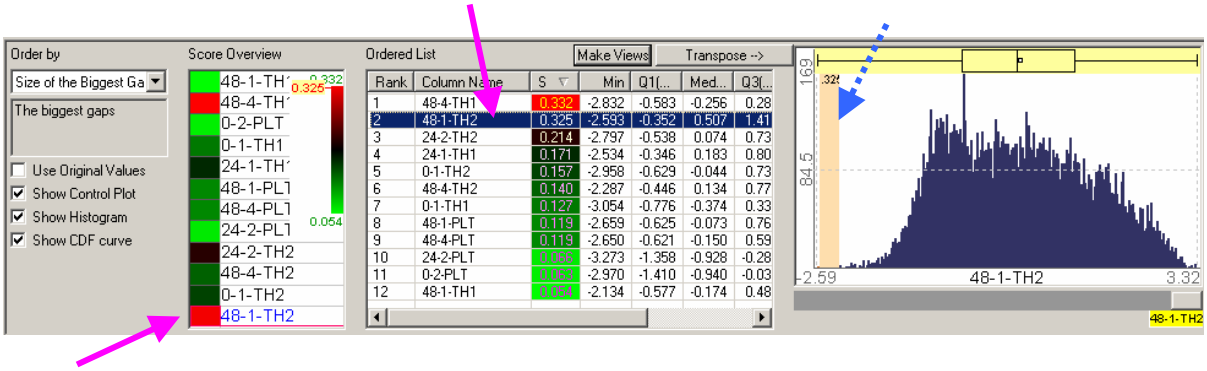


Figure 1 The biggest gap ranking result led P1 to make interesting discoveries. In this case the second biggest gap was especially interesting. It occurs in the histogram for 48-1-TH2 (purple arrow), and is shown as a peach color region (blue dotted arrow) on the left side of the histogram

At first P1 wrote down the probe set id and input this into NetAffix [1] in order to obtain ontological information. But this process could have been facilitated if P1 had used the gene ontology tab and annotation function available in HCE. Although P1 had been instructed in the use of the gene ontology tab, she did not use it when it would have been beneficial. After being reminded of the function, she tried it and found it useful and efficient.

P1 investigated the behavior of this probe set in other histograms using the histogram browser and discovered that the expression of this same probe set was consistently low in all TH2 samples (and progressively more so with time) and that it was consistently at a higher expression

level in TH1 and Platelet cells. The behavior of a probe set like this is of interest to this project because TH1 and TH2 cells have few unique cell markers, which makes it hard to identify and separate them from one another. So any gene that is very differentially regulated is of potential interest as a distinct cell marker and worthy of follow-up investigation. It is very important to have good cell markers for cell identification and separation because the balance of TH1 and TH2 cells is thought to influence the progression (recovery or fatality) of the sepsis patient.

3.2.2 Scatterplot Ordering

P1 tried all ranking criteria in the order that they appear in the combobox. With the very first ranking criterion, Pearson correlation coefficient, P1 noticed that samples of the same cell type were more highly correlated regardless of time point (Figure 2). This makes sense because the global pattern of gene expression would still be expected to be relatively cell specific and maintained from sample to sample. She also noted that there was a strong correlation between one of TH1 samples and Platelet samples (but not between the Platelet and TH2 samples). This is interesting in the context of other microarray analysis that was performed on this data set in GeneSpring (Silicon Genetics, Redwood City, CA) in which certain genes were identified that may be involved in Platelet regulation of the TH1/TH2 balance. This observation encourages further evaluation of the regulatory relationship between platelets and TH1 cells; this is a general trend but it may have been missed with other analysis tools.

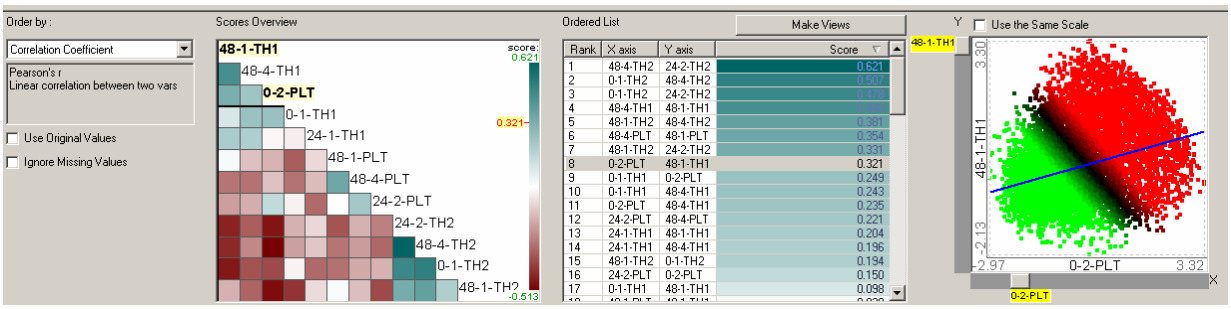


Figure 2 Scatterplot ordering result by correlation coefficient: high positive correlations in turquoise, high negative correlations in maroon

3.2.3 Discussion

This case study with P1 showed that HCE informed the researcher's overall analysis strategy and contributed to the analysis in a unique manner. First of all, HCE's unique framework using unsupervised clustering to enable researchers to decide which probe set interpretation method to choose for their Affymetrix projects [17] attracted P1 to start using HCE. While looking into the sample clustering result and the F-measure, users usually explored the histogram ordering tab to understand distributions of samples. Then with no instruction, users move on to the scatterplot ordering tab to understand relationships between samples. Of course, this natural work flow occurs more frequently as users become more proficient with HCE. Interactive coordination between the rank-by-feature framework and other displays such as the dendrogram and gene ontology views enables users to draw more specific conclusions.

Overall, P1 reported that: "There are several features that HCE offers that other programs do not with the most notable being the rank by feature functions. To my assessment, these tools allow a relatively speedy overview of the *shape* of one's data. I would therefore use these sorts of features at the beginning of my analysis to note any general trends that are taking place so that I can have those in mind as I execute my subsequent analyses." More specifically, P1 commented, "A great example of when this would have been helpful – I recently started analysis

on a data set processed by someone else; the data was already loaded onto GeneSpring etc. and as I was looking at specific lists of genes. It eventually became apparent that there was something strange going on with several of my time points (which was strange because all of the quality control data for the samples looked fine). When I loaded the data into HCE – this *strangeness* was immediately apparent - some of my disease samples were behaving much more similarly to the controls than to the other disease samples. I would have saved a large amount of time if this data set had been loaded onto HCE to begin with and I had been able to notice that these samples had strange trends and should be carefully evaluated."

Given all of the above, HCE adds some steps/perspectives to P1's analysis strategy rather than changing it all together. By far, P1's main analysis tools were dChip and GeneSpring, mostly because of their capability of comparing groups to find statistically significant differences in gene expression. P1 also liked GeneSpring's ability to load in experiment parameters and save large numbers of gene lists which can be compared across projects. However, through HCE's rank-by-feature framework and interactive visualization techniques, P1 found additional important information. P1 said she would definitely use HCE for future projects especially at the beginning of her analyses.

The data set used in this case study is still being evaluated - so it will be a little while before P1 publishes anything. At this point, P1 is following up on genes with specific behavior patterns that P1 hopes to confirm. P1 did actively use HCE to determine which signal interpretation algorithm was the most reliable for this analysis, and that should eventually be published in the methods section of upcoming papers.

3.3 FAMuSS Study Data Set

P2 is the principal statistician for the Center for Genetic Medicine, Children's National Medical Center. Most of the data analyses P2 performs are epidemiological in nature and includes large, multi-center genetic association studies. P2's everyday analysis tool was SAS, and P2 had almost no experience in using interactive visualization tools like HCE before this case study. We had two one-hour training sessions with P2. Since P2 is an expert in statistics, it was much easier to explain the rank-by-feature framework to P2 than to any other participant. Most of P2's data is collected prospectively, thus data exploration is a major part of P2's ongoing data analysis duties.

P2 loaded a multidimensional data set from the functional single nucleotide polymorphisms associated with muscle size and strength (FAMuSS) study [25]. FAMuSS study is a multicenter, NIH-funded program to examine the influence of gene polymorphisms on skeletal muscle size and strength before and after resistance exercise training. About one thousand men and women, age 18-40 year, were enrolled and trained their nondominant arm for 12 weeks. Skeletal muscle size and isometric and dynamic strength were measured before and after training. This data set has about 150 variables including anthropomorphic data, muscle strength data, and muscle, bone and fat size data. Some of the measurements were done for only a subset of participants, which means that there are many missing values (about 40%) in the data set.

Since this study was performed in an early stage of data analysis, most of the findings in this study were about quality of data sets and confirmation of expected relationships. As the data set becomes more complete, more interesting findings could be possible.

3.3.1 Histogram Ordering

P2 commented about the histogram ordering that "This feature is extremely useful to me as a statistician, mostly for data exploration. It allows me to look at the distributions and test

normality of all variables quickly and simultaneously. Additionally useful are the listings of outliers and numbers of unique values. Typically gaining this type of information using statistics packages is very time consuming, requiring an individual test and/or graph made for each variable."

P2 started to overview the clustering results on the dendrogram view after loading the data set as do most HCE users. Unlike microarray researchers, however, without spending much time examining clustering results, P2 tried the histogram ordering. Normality criterion first attracted P2, and P2 found that several variables such as baseline 1-RM (one repetition maximum) strength showed a bimodal distribution. It is important to know this because subsequent statistical analyses might be influenced by that.

By applying the biggest gap ranking and manually controlling the histogram browser, P2 could make a list of suspect data points including a subject with a BMI (body mass index) of 2.0 and a subject who has an exceptionally isometrically strong dominant arm. Follow-up examinations not only identified some data errors but also confirmed that some of the values were correct extreme ones. These findings of outliers are very important because it could lead to either development of a better analysis model or identification of interesting genes that caused the exception. The rank-by-feature framework enabled P2 to perform such important tasks more naturally and quickly.

3.3.2 Scatterplot Ordering

P2 summarized that "I find this feature one of the most useful to statistical analysis. By calculating scatterplots for every pair of variables, it not only allows the comparison of the plots of all continuous variables in a pair-wise fashion, but also allows simultaneous calculation of correlation coefficients and assessments of both linear and quadratic relationships. Obtaining this

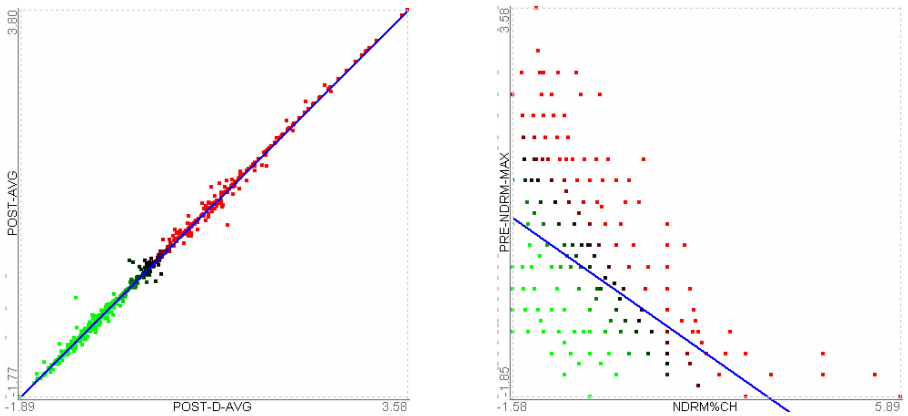
information from a statistics package again can be extremely time-consuming. I could save sometimes a hundred pages of SAS text output."

In the scatterplot ordering, the most interesting ranking criterion was "correlation coefficient" as it was for many other users. It turned out again that the linear correlation is one of the most interesting and important features that researchers want to detect as they start a multidimensional data analysis. At first, P2 tried to verify the trivial correlations in the data set. This task does not provide any new insight into the data set, but it is still important because researchers can confirm the validity of their data set. For example, a non-perfect correlation coefficient between baseline and post-exercise height allowed P2 to pick out individuals whose height was measured differently at the two time points.

P2 could also easily identify several strange perfect negative correlations between variables on the score overview (bright blue cells in Figure 4). After quickly checking the corresponding scatterplots on the scatterplot browser, P2 could easily conclude that those perfect negative correlations were due to missing values. All those scatterplots actually have only one valid item and all other items are missing values. Problems caused by missing values led us to improve the rank-by-feature framework in a way that ranking results could be less susceptible to missing values, which will be discussed later in this section.

P2 could easily find groups of variables that have strong positive correlations. The score overview in Figure 4 shows triangular or rectangular red areas, which represent that corresponding variables are highly correlated (one example at Figure 3(a)). Those correlations include correlation between baseline and post-exercise measurements of 1-RM strength, isometric strength, and etc.

An interesting weak negative correlation between NDRM%CH (% change in 1-RM strength of non-dominant arm) and pre-NDRM-max (pre one repetition max of non-dominant arm) shown in Figure 3(b) was also detected on the score overview. This correlation might indicate that 1-RM strength of non-dominant arm improves less after 12 weeks exercise as the baseline 1-RM max is bigger. Simply speaking, 12 weeks exercise could make more positive changes to people who have a relatively weak arm.



(a) (b)

Figure 3 Selected scatterplot ordering results with FAMuSS Study data set

3.3.3 Discussion

Overall, P2 was impressed by interactive visual feedback of HCE. HCE has been most useful for its efficient visualization ability and calculation of basic statistics. Since P2 had not used the clustering feature before, she focused on the rank-by-feature framework that she thought were extremely useful to her as a statistician for data exploration. However, P2 also tried other features such as the color mosaic view and profile search, and found them also useful to see the magnitude of missing data and to quickly pick out data points that seem unusual.

P2 recommended a list of statistical tests as ranking criteria that she wanted to have in the future version of HCE, which includes Student t-test, ANOVA, Chi square, and some non-

parametric tests. We considered implementing these, but a more efficient way to add these ranking functions in future versions of HCE is to utilize implementations in other packages such as R, SAS, and Matlab. The linkage to those packages could greatly improve the usefulness of the rank-by-feature framework and HCE.

Since missing values were all set to 0 then for the rank-by-feature framework, ranking results involving line or curve fittings could be distorted by the missing values as shown in the scatterplot at the bottom right corner of Figure 4, where the regression line is dragged down significantly due to many missing values for the Y-axis. To solve this problem, we implemented a checkbox to enable users to exclude the missing values from the ranking function evaluation. This option significantly improved the ranking results for this case study data set. For example, the fitting result for the same variable pairs shown in Figure 4 was significantly improved by excluding missing values from the ranking function evaluation in Figure 5(b). Compared to the score overview in Figure 4, the ranking result by the correlation coefficient criterion was also significantly improved after excluding the missing values (Figure 5(a)).

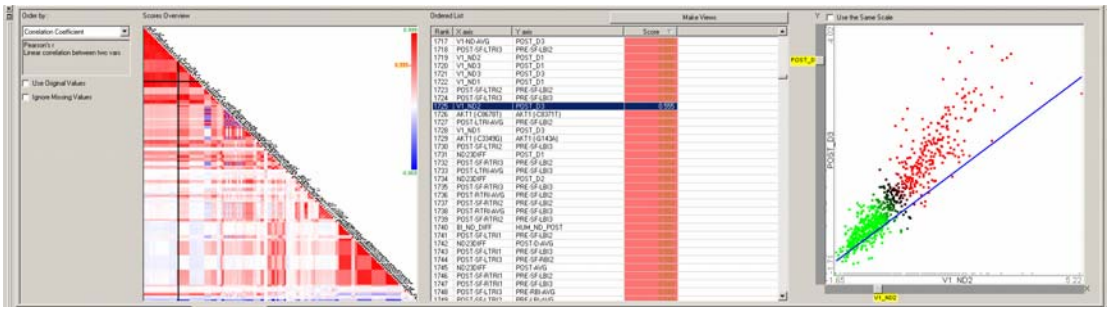
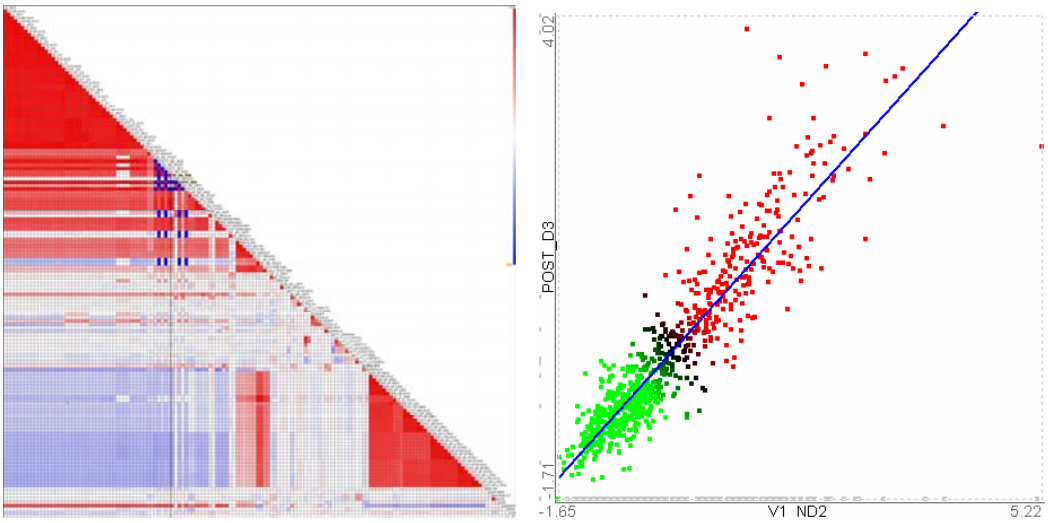


Figure 4 FAMuSS Study data set in HCE



(a) (b)

Figure 5 An improved score overview (a) and fitting result (b) with missing values excluded

One important issue in this case study was the problem of dealing with a large number of variables. On a common monitor with resolution of 1280x1024, the score overview is so crowded that variable names are barely readable. A high resolution monitor (e.g., 3840x2400) could reduce this problem. A zooming, filtering, or grouping control for the rank-by-feature framework would be useful addition to cope with large numbers of variables.

P2 used HCE to do most of her data exploration at the start of analysis, so HCE actually contributed to all of the papers that have come out of the FAMuSS Study. The most significant contribution was made in discovering a strong association between AKT1 haplotypes and body composition in males, which is under review for *the American Journal of Human Genetics*.

3.4 Aerosols, Clouds, and Precipitation

A researcher (P3) in the meteorology department at the University of Maryland was interested in using HCE for his research projects. After two demonstration sessions, P3 was convinced that his research could benefit from HCE, and agreed to participate in the case study. P3 said that data clustering is not necessarily required in his research field, but he often needs to stratify the

data. P3 had mostly used spreadsheet software such as Excel and Sigmaplot [23] to view correlation and distribution for some variables of importance. P3 had also been learning and using IDL (Interactive Data Language) [14], which is a programming environment similar to MATLAB [24] and popular in the meteorology field.

The data set for this case study was an in situ aerosol profiling data, which has 2829 rows (time) and 23 columns (measurements). The variables used for the analysis include amount and size of aerosols, and various meteorological conditions relevant to aerosols – cloud amount, wind, relative humidity, etc. P3's intended usage of HCE was to classify aerosols according to meteorological conditions and to identify which conditions result in stronger relationships among the variables representing aerosol load and properties.

3.4.1 Histogram Ordering

P3 used the histogram ordering when he investigated the data set for the first time. He tried all the ranking criteria to find the gap size ranking and the normality ranking most interesting. From the normality ranking result, P3 could preattentively notice that AOT_670 (Aerosol optical depth measured at the wavelength of 670nm) showed the least normal distribution. On the histogram browser he realized that it has a bimodal distribution, and it also has several distinctive outliers, which were also easily noticeable in the ranking result by the biggest gap size.

Unlike other case study participants, P3 wanted to move on to the scatterplot ordering after quickly trying the histogram ordering. This was in part because he was much more interested in pair-wise relationships than individual distributions. P3 was also distinctive in the way he used HCE. He was interested in finding relationships using all data items and also with only some subsets of items such as those falling into a cluster of items. He loved to see the coordination between the dendrogram view and the rank-by-feature interface. When he examined a ranking

result, he selected many clusters one by one in the dendrogram view and saw how the items in the cluster were distributed in a histogram or in a scatterplot.

3.4.2 Scatterplot Ordering

A couple of scatterplots (e.g. ‘wind speed’ vs. ‘wind direction’ and ‘aerosol optical depth’ vs. ‘aerosol concentration number’) from the correlation coefficient ranking attracted P3’s attention. P3 would like to investigate two scatterplots at the same time by highlighting items with one wind direction and then highlighting others with the opposite wind direction. P3 found two well-defined groups on both scatterplots in terms of their wind-direction.

P3 unexpectedly saw a relationship between two variables, which was never examined before. That was the quadracity between cloud fractions computed at two different circumsolar areas (Figure 6). Instead of being satisfied by the finding, P3 used the dendrogram view to determine which cluster contributed to the quadratic relationship. P3 identified two clusters - one with well-defined quadracity (B in Figure 6) and the other with break-down of such quadracity (A in Figure 6). P3 did not stop here; instead he examined other relationships among aerosol-related parameters for the selected two clusters to check if it makes any difference.

At the first weekly meeting where he explained his finding of the quadratic relationship, P3 complained that he could not see more than one scatterplot at the same time. Even though we had explained how to do it at the demonstration sessions, he forgot how to do it. Being reminded of it at the next meeting, he could investigate relationships much more efficiently by looking at two or more scatterplots at the same time. P3 finally found another interesting feature: the well-defined quadracity was involved in relatively low water vapor amount regardless of aerosol number concentration, whereas the break-down of quadracity was involved in low aerosol number

concentration regardless of water vapor amount (two scatterplots at the bottom in Figure 6). This interesting feature might improve the underlying model later after further investigation.

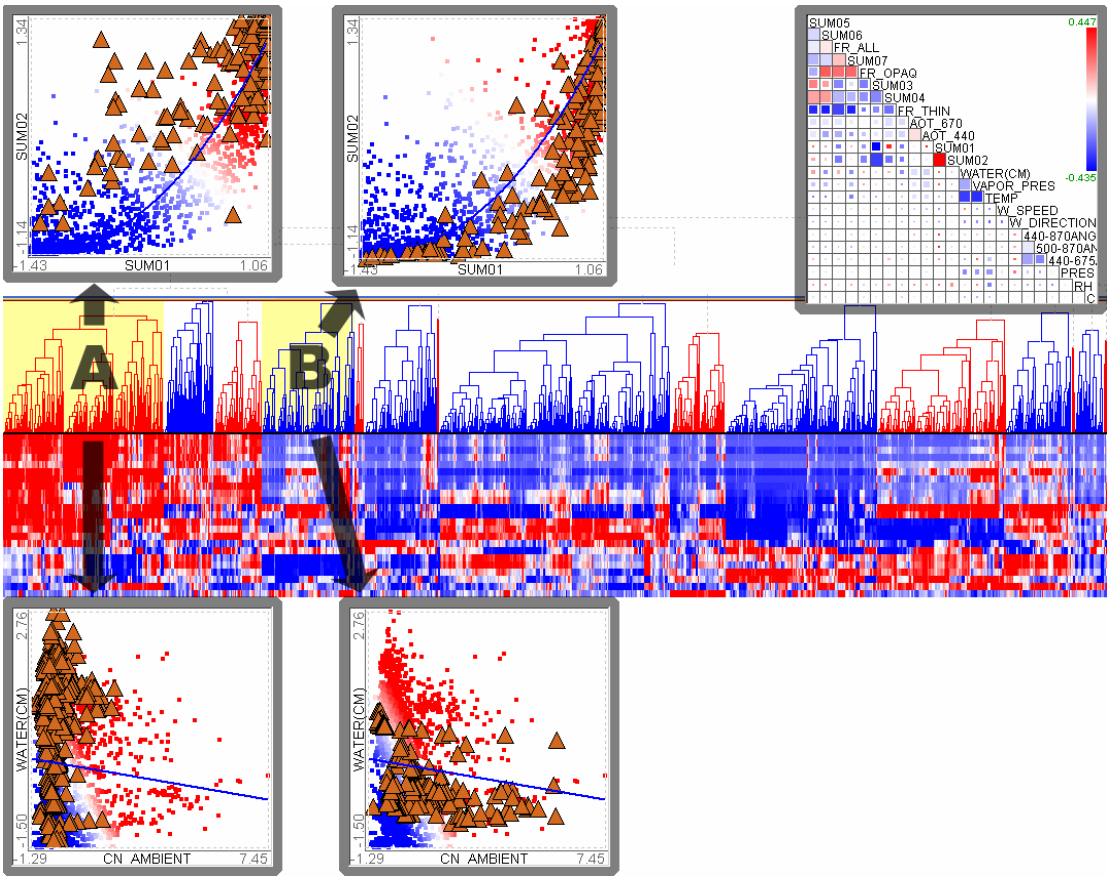


Figure 6 High quadracity found in the aerosol data set. Score overview is at top right corner, where a big bright red cell occurs for SUM01 and SUM02. Size coding is by complement of least square error and color coding by the score (coefficient of the highest term). Two scatterplots at the top show the quadracity between SUM01 and SUM02. The left scatterplot highlights items in cluster A, and right scatterplot highlights items in cluster B. The two scatterplots at the bottom show distinctive distributions of two clusters on a 2D projection (CN_AMBIENT vs. WATER).

3.4.3 Discussion

Overall, P3 was thrilled by the interactivity and visual feedback of HCE, and was very interested in using interactive multiple views coordination. P3 commented that “The main utility of HCE in my study is to quickly view data histograms, relationships (e.g., correlation) between

variables, and to stratify the data, if necessary. Since HCE does the jobs all at once, it is a very convenient tool for data *quick-look*."

P3 suggested adding scaling functions to the rank-by-feature framework to effectively deal with various types of units and distributions of variables. Users could scale each variable in the histogram ordering before ranking, and the scaling result could affect the ranking in the scatterplot ordering. Considering that many other users had also suggested the similar idea, this functionality could improve usefulness of the rank-by-feature framework as well as other HCE tools.

At the first demonstration session with P3, he asked for a function to customize color mapping in the score overview. At the time, HCE only used green and red color coding by default, and users could not customize it. He preferred a red-blue color scheme intermediated by white color, which has been widely used in the meteorology research field. We accepted this request and implemented it in the next version of HCE, which was used for this case study. Another suggestion by P3 related to color use in HCE is the function of changing background color for each view in HCE, especially for scatterplot views.

This case study also identified a potential future implementation possibility. Most multiple views coordination systems maintain only one set of selected items which are highlighted in all coordinated views. If multiple sets of selected items are allowed, it could improve cognition of important patterns in some cases. For example, if users could select two clusters and color each cluster differently in Figure 6, users might see the quadratic relationship more clearly in a single scatterplot view or two separate views. Furthermore, if the intersection of sets of selected items is colored differently when the sets could be non-disjoint, users could visually scrutinize the interaction among those sets.

A follow-up investigation into the quadracity between SUM01 (cloud fraction for circumsolar region within angular distance between 10-30 degree from the direction of the solar beam) and SUM02 (the same for 10-40 degree) enabled P3 to figure out a possible case of it, which was related to the cloud detection algorithm that was used for the cloud amount measurement. He hypothesized that the cloud detection algorithm might overestimate the amount of clouds at the inner circumsolar areas (SUM01) due to the difficulty in cloud detection near the sun. This hypothesis needs to be validated through further investigations. If the hypothesis is accepted, it might contribute to the development of a better cloud detection algorithm.

3.5 Conclusion

Month-long case studies with motivated users gave us a chance to look closely at how HCE and the rank-by-feature framework are used for research projects. It became clear that HCE and the rank-by-feature framework enable users to quickly examine their data sets in ways that pleased our participants. The GRID principles seemed to be naturally applied by most participants as if the principles had been accepted for a long time. Interactive visual examinations often led to the identification of important unexpected patterns in the data set, which is important for data verification and hypothesis generation.

Even though HCE is more stable than other research prototypes freely available, it had crashed several times over the course of the case studies. Participants' understanding and willingness to accept these problems enabled case studies to finish successfully with invaluable suggestions and improvements. Regular meetings and prompt email communication were important means by which we could make the participants feel as if we were research partners rather than merely using them as test subjects. One of most difficult parts of these kinds of case studies is that the developer of the tool needs to spend ample time to understand the data set and

the underlying research problems that participants are interested in. Without such understanding, it is not easy to make participants think of the experimenter as a research partner. Another difficult part was that sometimes a participant might forget what had been done in earlier meetings. This is in part because the interval between meetings, usually a week, was too long. A better option could be a one-week intensive case study. However, this option also has its shortcomings. Participants' research might be distracted by frequent meetings, and important design suggestions from participants could not be promptly incorporated into the tool and the case study itself.

Overall, although there were a couple of cases of early termination, case studies showed the efficacy of HCE and the underlying GRID principles for the analyses of multidimensional data. Invaluable suggestions for improvement were also made by participants, which include: (1) color coding customization, (2) missing value handling in ranking functions, (3) scaling of each variable, (4) multiple selection sets, (5) potential ranking criteria including various important statistical tests, and (6) linkage to external statistical tools.

4 HCE User Survey via Email

HCE has been freely distributed on the web at www.cs.umd.edu/hcil/hce for research or academic purposes. As of February 2005, about 2451 downloads have been logged in the download log since we opened up the download page in April 2002. More people download HCE as newer versions are released (196 in 2002, 822 in 2003, 1229 in 2004, and 1600 expected in 2005).

In spite of the complex statistical analyses, users from around the world downloaded HCE. Its most popular users are biologists doing microarray data analysis of gene expression data, but other interesting users include social scientists, defense or security agencies, environmental or

financial analysts. It is used in various educational settings, business data analysis, and has been licensed to a biotech company at New Zealand.

When users download HCE from the HCE homepage, they are asked to fill in the registration form. There is an optional field where about one-third of users wrote down their intended usage of HCE. Encouraged by this and many email inquiries from HCE users, we decided to conduct an email user survey on the usage of the rank-by-feature framework and HCE. After removing duplicated email addresses and roughly filtering invalid email addresses, we sent out the user survey questionnaire to about 1500 email addresses. The questionnaire consists of 13 questions regarding HCE usage in general and the rank-by-feature framework. Almost 40% of user survey emails were undelivered due to various reasons such as invalid email address and blocking by spam filters. Finally, 83 users replied, which is around 9% of all users from whom the survey email was at least not bounced. Among the 83 users, 26 users did not answer a majority of questions because they did not actually use HCE or just tried it for curiosity. Thus, this section summarizes the answers of 57 users.

4.1 HCE: Overall

Most of the users are biologists, computer scientists, and statisticians, but physicists, business managers, sociologists, geographers, and medical doctors are also users. Microarray data analysis and clustering analysis are the most popular uses of HCE. HCE is also used as a teaching tool for information visualization and data mining classes.

A large portion of users run HCE with their data set just to quickly examine a hierarchical clustering result (*How often did you use HCE when you used it most intensively?*: 22 for once a month, 10 for once a week, 7 for once a day, and 15 for many times a day). Sometimes they just get a screen grab of the dendrogram. Interestingly, some users use HCE many times a day to

explore their data using various features in HCE. Most of these active users tend to think that HCE significantly improved the way they analyze data sets while most less active users (once a month) think HCE had a modest impact. More users tried HCE with fairly large data sets than with small data sets (*What is the maximum number of rows in data sets that you have loaded in HCE?*: 8 for “less than 100”, 11 for “less than 1000”, 11 for “less than 10000”, and 19 for “more then 10000”). This is partially because many users tried to analyze microarray data sets where there are commonly more than 10,000 rows, or sometimes around 40,000 rows. Because the number of columns does not significantly affect the performance of most features in HCE, we did not ask about the number of columns, but it is mostly from 10 through 150.

Since HCE had become known to most users as a cluster visualization tool, they used the dendrogram and color mosaic feature (*Which features have you used?*: 49 for “dendrogram”, 25 for “histogram ordering”, 25 “for scatterplot ordering”, 25 for “tabular view”, 22 for “profile search”, and 7 for “gene ontology”). Since, our tabular view uses a list view control that improves on the standard Windows version [20], it was pleasing to find that many users rated it helpful for data exploration. The rank-by-feature framework (histogram and scatterplot ordering) were also used frequently by many users, thereby supporting our claims. The gene ontology view is only useful to molecular biologists who are interested in gene ontology, so it is used by the smallest number of users. Generalizing the gene ontology view to other hierarchical knowledge structures might improve its usefulness (e.g. to sociologists or business analysts).

4.2 Rank-by-Feature Framework

More users said it was easier (very easy or somewhat easy) to use the histogram ordering (53%) than the scatterplot ordering (46%). This might be in part because relationships between variables are more difficult to appreciate than each individual variable alone. According to users’

additional comments, it seems clear that users try the histogram ordering first and then the scatterplot ordering, which is consistent with the GRID principles.

The ranking criteria are more evenly useful in the histogram ordering than in the scatterplot ordering (Figure 7 & Figure 8). Ranking criteria in the histogram ordering seem to be easier to understand than those in the scatterplot. The least square error and quadracity criteria in the scatterplot ordering are the most difficult for users to understand. Explanations of ranking criteria shown in the rank-by-feature framework and HCE homepage might be too short to make users understand the ranking criteria. Context-sensitive help or an online help page could encourage users to use such difficult but sometimes useful ranking criteria.

In both orderings, the first ranking criterion, normality for the histogram ordering and correlation coefficient for the scatterplot ordering, is most popularly used. Considering that average HCE users are professionals who have some knowledge of statistics, the implication of the normality test may be well understood by most users. Other ranking criteria in the histogram ordering are also almost straightforward. “The size of the biggest gap” ranking criterion is a novel concept, so it is least utilized even though the idea is simple. As shown in case studies, once users get the idea of the gap, it becomes a very useful ranking criterion for outlier detection.

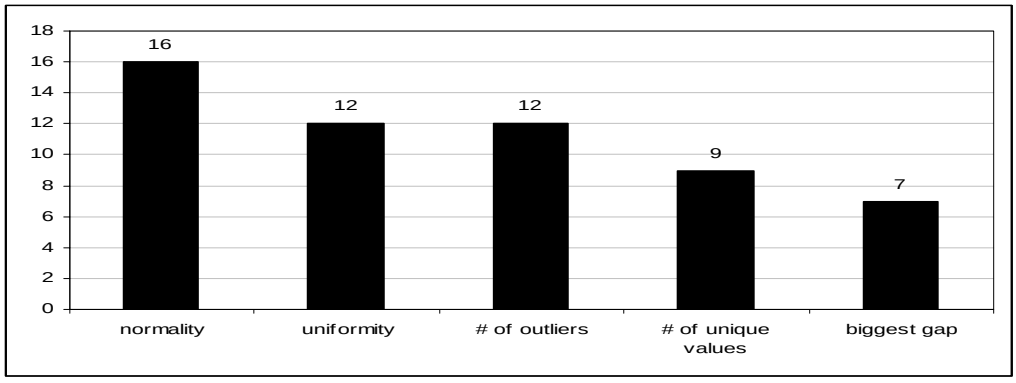


Figure 7 What are the most useful ranking criteria in the histogram ordering?

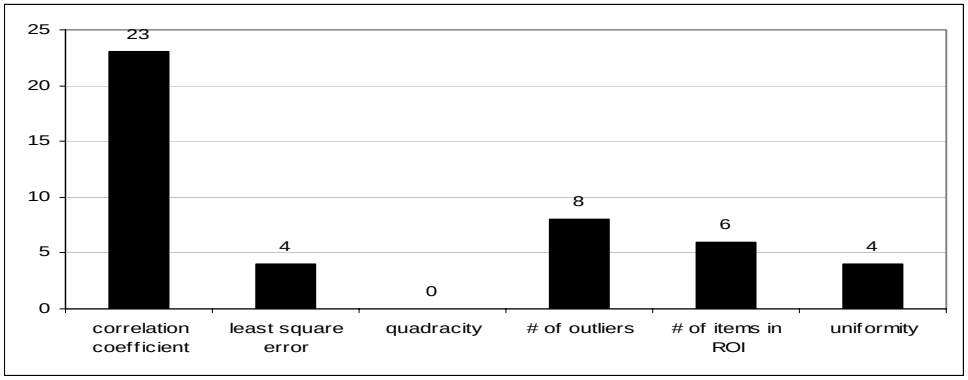


Figure 8 What are the most useful ranking criteria in the scatterplot ordering?

Correlation is an important and well known linear association between two continuous variables. Thus, after users decided to try the scatterplot ordering, they would at least try this first ranking criterion, correlation coefficient. Most users find the score overview is useful to examine correlations between variables. A participant commented that the complete overview of all possible pair-wise relationships prevent potential problems caused by missing some important relationships by chance. Even though uniformity and the number of outliers are 2D versions of the same ranking criteria in the histogram ordering, users seemed to have some difficulty in applying them to 2D relationships. No participant voted for the quadracity criterion. Although a case study participant (P3) found it useful, more work could improves its acceptance.

4.3 Discussion

About 96% of users said that HCE improved the way they analyze their data sets at least a little bit, and about 73 % of those users felt that HCE at least somewhat significantly improved their analysis practices (13 for “significantly”, 20 for “somewhat significantly”, 12 for “a little bit”, and 2 for “not at all”). 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!"

Users' additional comments indicate that interactive visual presentations and sustainable robustness of HCE get credit for that. Together with appreciation for making HCE available, users suggested several improvements: (1) some evaluation measures for unsupervised clustering results, (2) more clustering algorithms or other projection techniques such as SOM and PCA, and (3) more elaborate import/export/ print/save functions.

A few users also expressed their concern over the point that some ranking criteria are difficult to understand without deep statistical backgrounds. This is actually a very difficult problem to address appropriately. Even after a thorough live demonstration session, a couple of users still have a difficulty in understanding some ranking criteria. Detailed tutorials could help users go through if they are motivated. Otherwise it is not a general solution. This problem is related to whether a tool is for a general audience or for specialized users. The current version of HCE requires some statistical knowledge, which makes it a more sophisticated tool.

This user survey certainly had its limitations. First, even though users' responses to the survey email were voluntary, there was still a danger that users who had been disappointed with HCE were less likely to participate. If we had randomly selected participants, the result might have been different from the current result. However, it would have been difficult to compel the randomly selected users to participate in the survey. Second, the number of participants was limited. If the survey had been conducted via a web page instead of emails, the turnout might have been better due to the better-preserved anonymity. Third, a problem related to the design of the questionnaire meant that several respondents made only one selection for multiple-selection questions.

In spite of the limitations, this user survey showed the usefulness of HCE and the rank-by-feature framework in terms of improving the way users analyze their data. The GRID principles seemed to be implicitly observed, but more work is necessary to encourage more users to smoothly advance from 1D study to 2D study. More training materials and context sensitive help are necessary to help users understand the utility and implication of ranking criteria.

5 Conclusion

This paper culminates our three year effort in building, applying, evaluating, and refining a powerful knowledge discovery tool for multi-variate and high-dimensional data. We believe that the guiding GRID principles and especially the rank-by-feature framework can be useful to designers of other information visualization tools. Since it is difficult to conduct controlled experiments on complex tools that require substantial training and changes to analytic processes, we conducted three longitudinal case studies and an email user survey. Our case studies included three participants from different research fields who are accustomed to their distinctive analysis practices. The email user survey makes it possible to get a more general feedback from a variety of users who applied HCE in their natural working environment conducting their tasks. We hope that these contextual evaluations will contribute to (1) understanding how exploratory strategies such as the GRID principles and the rank-by-feature framework can influence design, (2) attracting new users to information visualization tools such as HCE and (3) encouraging knowledge discovery tool designers to adopt similar evaluation approaches.

Acknowledgement: This work was supported by N01 NS-1-2339 from the NIH and by the National Science Foundation under Grant No. EIA 0129978. We also thank Ben Bederson, Catherine Plaisant, and other HCIL members for constructive suggestions during this work.

References

[1] Affymetrix, NetAffix, https://www.affymetrix.com/analysis/netaffx/index.affx

[2] P. L. Bollyky and S. B. Wilson, "CD1d-restricted T-cell subsets and dendritic cell function in autoimmunity," *Immunology and Cell Biology*, vol. 82, pp. 307-314, 2004.

[3] W. H. Boylston, A. Gerstner, J. H. DeFord, M. Madsen, K. Flurkey, D. E. Harrison, and J. Papaconstantinou, "Altered cholesterologenic and lipogenic transcriptional profile in livers of aging Snell dwarf (Pit1dw/dwJ) mice," *Aging Cell*, vol. 3, pp. 283-296, 2004.

[4] C. Chen and M. P. Czerwinski, "Empirical evaluation of information visualizations: an introduction," *International Journal of Human-Computer Studies*, vol. 53, pp. 631-635, 2000.

[5] C. Chen and Y. Yu, "Empirical studies of information visualization: a meta-analysis," *International Journal of Human-Computer Studies*, vol. 53, pp. 851-866, 2000.

[6] S. Cluzet, C. Torregrosa, C. Jacquet, C. Lafitte, J. Fournier, L. Mercier, S. Salamagne, X. Briand, M. T. Esquerre-Tugaye, and B. Dumas, "Gene expression profiling and protection of Medicago truncatula against a fungal infection in response to an elicitor from green algae Ulva spp," *Plant, Cell and Environment*, vol. 27, pp. 917-928, 2004.

[7] G. K. Geiss, M. Salvatore, T. M. Tumpey, V. S. Carter, X. Wang, C. F. Basler, J. K. Taubenberger, R. E. Bumgarner, P. Palese, M. G. Katze, and A. Garcia-Sastre, "Cellular transcriptional profiling in influenza A virus-infected lung epithelial cells: The role of the nonstructural NS1 protein in the evasion of the host innate defense and its potential contribution to pandemic influenza," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, pp. 10736-10741, 2002.

[8] V. Gonzales and A. Kobsa, "A Workplace Study of the Adoption of Information Visualization Systems," in *Proc. I-KNOW'03: Third International Conf. Knowledge Management*, 2003, pp. 92-102.

[9] A. M. Graziano and M. L. Raulin, *Research Methods: A Process of Inquiry*, 5 ed: Allyn & Bacon, 2004.

[10] C. Li and W. H. Wong, "Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, pp. 31-36, 2001.

[11] H. Lieberman, The Tyranny of Evaluation, http://web.media.mit.edu/~lieber/Misc/Tyranny-Evaluation.html

[12] E. Paux, V. Carocha, C. Marques, A. Mendes de Sousa, N. Borralho, P. Sivadon, and J. Grima-Pettenati, "Transcript profiling of Eucalyptus xylem genes during tension wood formation," *New Phytologist*, vol. 167, pp. 89-100, 2005.

[13] C. Plaisant, "The challenge of information visualization evaluation," in *Proceedings of the working conference on Advanced visual interfaces*. Gallipoli, Italy: ACM Press, 2004, pp. 109-116.

[14] Research Systems Inc., Interactive Data Language, http://www.rsinc.com/idl/

[15] P. Saraiya, C. North, and K. Duca, "An Insight-Based Methodology for Evaluating Bioinformatics Visualizations," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 11, pp. 443 - 456, 2005.

[16] J. Scheidtmann, A. Frantzen, G. Frenzer, and W. F. Maier, "A combinatorial technique for the search of solid state gas sensor materials," *Measurement Science and Technology*, vol. 16, pp. 119, 2005.

[17] J. Seo, M. Bakay, Y.-W. Chen, S. Hilmer, B. Shneiderman, and E. P. Hoffman, "Interactively optimizing signal-to-noise ratios in expression profiling: project-specific algorithm selection and detection p-value weighting in Affymetrix microarrays," *Bioinformatics*, vol. 20, pp. 2534-2544, 2004.
[18] J. Seo, M. Bakay, Z. Po, C. Yi-Wen, P. Clarkson, B. Shneiderman, and E. P. Hoffman, "Interactive color mosaic and dendrogram displays for signal/noise optimization in microarray data analysis," in *Proceedings of IEEE International Conference on Multimedia and Expo*, 2003, pp. III-461~III-464.
[19] J. Seo and B. Shneiderman, "Interactively exploring hierarchical clustering results," *IEEE Computer*, vol. 35, pp. 80 - 86, 2002.
[20] J. Seo and B. Shneiderman, "A Knowledge Integration Framework for Information Visualization," *Lecture Notes in Computer Science*, vol. 3379, pp. 207 - 220, 2005.
[21] J. Seo and B. Shneiderman, "A rank-by-feature framework for interactive exploration of multidimensional data," *Information Visualization*, vol. 4, pp. 99-113, 2005.
[22] Silicon Genetics, GeneSpring, http://www.silicongenetics.com/cgi/SiG.cgi/Products/GeneSpring/index.smf
[23] Systat Software Inc., SigmaPlot, http://www.systat.com/products/SigmaPlot/
[24] The MathWorks, MATLAB, http://www.mathworks.com/products/matlab/
[25] P. D. Thompson, N. Moyna, R. Seip, T. Price, P. Clarkson, T. Angelopoulos, P. Gordon, L. Pescatello, P. Visich, R. Zoeller, J. M. Devaney, H. Gordish, S. Bilbie, and E. P. Hoffman, "Functional polymorphisms associated with human muscle size and strength.," *Medicine & Science in Sports & Exercise*, vol. 36, pp. 1132-1139, 2004.
[26] J.-M. Tsai, H.-C. Wang, J.-H. Leu, H.-H. Hsiao, A. H. J. Wang, G.-H. Kou, and C.-F. Lo, "Genomic and Proteomic Analysis of Thirty-Nine Structural Proteins of Shrimp White Spot Syndrome Virus," *Journal of Virology*, vol. 78, pp. 11360-11370, 2004.
[27] P. Zhao, J. Seo, Z. Wang, Y. Wang, B. Shneiderman, and E. P. Hoffman, "In vivo filtering of in vitro expression data reveals MyoD targets," *Comptes Rendus Biologies*, vol. 326, pp. 1049, 2003.