Metrika (2000) 51: 39–51

Metrika

Interactive visualization of hierarchical clusters using MDS and MST

Sung-Soo Kim[1], **Sunhee Kwon**[2] **and Dianne Cook**[2]

[1] Department of Information Statistics, Korea National Open University, 169 Dongsung-Dong Chongro-Gu, Seoul 110-791, S. Korea (e-mail: sskim@mail.knou.ac.kr)
[2] Dept. of Statistics, ISU, 325 Snedecor Hall, Ames, IA 50011-1210, USA
(e-mail: dicook@iastate.edu)

Abstract. In this paper, we discuss interactively visualizing hierarchical clustering using multidimensional scaling (MDS) and the minimal spanning tree (MST). We can examine the sequential process leading to agglomerative or divisive hierarchical clustering, compare the different agglomerative methods, and detect influential observations better than is possible with dendrograms.

Key words. Hierarchical cluster analysis, multidimensional scaling (MDS), minimal spanning tree (MST), interactive visualization, statistical graphics, data mining, grand tour, dynamic graphics.

1 Introduction

Clustering is a valuable method for understanding the complex nature of multivariate relationships, and it is widely used in taxonomy and pattern recognition. It is enjoying a resurgence in popularity in the context of data mining. Cluster analysis is a procedure for grouping individuals or objects into similarity groups, without prior knowledge of groups. It is an exploratory tool. In general, the methods are divided into two categories: hierarchical and non-hierarchical. We focus on hierarchical methods.

In hierarchical cluster analysis the algorithms begin (1) with all objects in a single large cluster and proceed to sequentially divide them into smaller clusters, or equivalently, (2) with all the objects as individual clusters and proceed to sequentially fuse or agglomerate them into larger clusters, based on the interpoint distances. More emphasis has been placed on the rules governing the splitting or fusing process than on the adequacy with which each

The implemented S-plus and JAVA source programs, and associated documentation can be obtained from Web site: www.public.iastate.edu/~dicook/papers/Metrika/paper.html.

cluster accurately represents the objects in the measurement space (Zupan, 1982). Graphical representation of the clusters, consequently, provides visual grouping information, and plays a complementary role to the numerical algorithms in cluster analysis.

In hierarchical cluster analysis, the clustering process can be summarized diagrammatically in tree form, i.e. a dendrogram. Using a dendrogram, we can see the sequence of successive fusions or divisions that occur in the clustering process. For example, in Figure 3 following downwards from the top to the bottom of the dendrogram we can get the divisive process of clustering, while following upwards from the bottom to the top we can get the agglomerative process. Different agglomerative methods can produce radically different dendrograms, and a single observation can dramatically affect the dendrogram. Essentially the dendrogram is good for displaying sequential steps in the hierarchical algorithms, but it lacks the context of the problem, i.e., relative positions of the points and their interpoint distances. For these reasons, the dendrogram is less helpful for comparing methods and detecting influential observations. Yet, these are important parts of a cluster analysis. Because cluster analysis is inherently exploratory, it is important to examine the results produced by different methods, and assess the impact of excluding certain observations. To extract this type of information different graphical representations are drawn: multidimensional scaling (MDS) views, profile plots, stars, Chernoff faces, Andrews curves and scatterplots. Adding interaction and motion to these graphical displays greatly enhances the exploratory process.

Buja, Cook and Swayne (1996) discuss an interactive system where the dendrogram is dynamically linked to a grand tour scatterplot display. The grand tour (Asimov, 1985) exposes clustering through motion of points as they are spun through high-dimensional euclidean space. Points that are "close" in the data space will have similar motion patterns. The dendrogram is also overlaid on the data as it moves in the grand tour. This helps in understanding the agglomerative process in terms of interpoint distance and can assist in detecting influential observations. Also, Swayne, Buja and Hubbell (1991) describe interactively "cutting" the dendrogram in an S-Plus plot which colors the corresponding observations according to the resulting cluster solution in scatterplot views in XGobi (Swayne, Cook, and Buja, 1998). Now, a more common approach to graphically representing the individuals is to use MDS to find a low-dimensional representation of the data that closely preserves the cluster structure. We discuss this approach further in this paper, and also discuss overlaying a minimal spanning tree (MST) rather than a dendrogram. The MST provides strong visual cues for unraveling cluster structure in high-dimensional data.

This paper discusses using MST with MDS for interactive visualization of hierarchical clustering. We demonstrate comparing agglomerative methods (single, complete, average, centroid, median, Ward) and detecting influential observations. The work is implemented in S-Plus, and some of it is implemented in a prototype JAVA program.

2 MDS and MST

Multidimensional scaling (MDS) is a method that provides a low-dimensional visual representation of points that preserves their relative positions in high-dimensions, using a dissimilarity matrix or rank orderings. Because MDS is a way of representing interpoint distances it can naturally be used for visually identifying clusters. It is commonly used for this purpose. An interesting feature of MDS is that it depends only on the dissimilarity matrix, and as a consequence it is useful even in situations where a raw data matrix is nonexistent. See Buja, Swayne and Littman (1998) for an example of an interactive system for generating MDS representations, and Bienfait and Gasteiger (1997) for an approach to visually describing the error in MDS representations. Note also that, MDS is a very computationally intensive procedure and is only feasibly computed in real-time for small data sets, and for larger data sets the MDS representation needs to be computed off-line. The MDS representation is independent of the hierarchical cluster analysis, so alone it is of little help to understand the agglomerative process. One needs to connect the idea of the dendrogram to the MDS representation.

An approach to solving this problem, along the lines of Buja, Cook and Swayne, is to overlay a representation of the dendrogram on the MDS plot. Here, an interesting point to note is that, the dendrogram is essentially the MST when the agglomerative method is single linkage. It is a very neat representation of the dendrogram in this case. The MST is a tree which completely connects all the objects without any closed loop and minimizes the sum of the edge lengths of the tree. Given n points, a tree spanning these points is any set of edges joining the pairs of points such that there are no closed loops, each point is visited by at least one edge, and the tree is connected (Gower and Ross, 1969). When a set of n points and $n(n-1)/2$ possible edges are given, a minimal spanning tree (MST) is a spanning tree for which the sum of edge lengths is smallest. MST is closely related to single linkage cluster analysis. Since single linkage joins clusters by the shortest length between them, single linkage cluster can be constructed from the MST. So a dendrogram for single linkage cluster can be drawn directly from the MST (see Gower and Ross, 1969). The MST has been used for identifying influential observations by Jolliffe et al. (1995), and for highlighting the inaccuracies present in the low-dimensional MDS representations of high-dimensional data by Gordon (1981), Krzanowski (1988), Bienfait and Gasteiger, (1997). The MST provides valuable visual cues to cluster structure when used in conjunction with a scatterplot.

The structure overlaid on the plot can be adapted from the strict MST to be a useful representation of the dendrogram even with other agglomerative methods. For example, within the groups use MST to visualize the nature of the interpoint connectedness here, and between groups connect the elements using a representation that matches the agglomerative method. We will simply connect the closest elements between groups by a line. This provides sufficient information for us to compare methods.

3 Adding interaction

We provide MDS representations overlaid by the MST in an interactive setting, allowing the user to change the number of final clusters to examine the agglomerative or divisive sequence, compare different agglomerative methods, and the influence of particular objects on the final solution. We introduce an example to demonstrate the methods: villages data introduced by Morgan and Shaw (1982), and used again by Seber (1984) and Jolliffe et al. (1995).

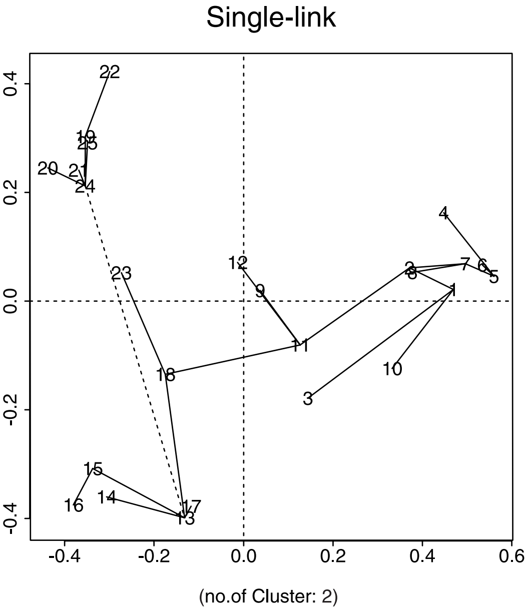

Fig. 1. Single linkage: 2 clusters

The villages data comes as a similarity matrix, measuring the frequency that the same word is used for a range of 60 items amongst 25 villages. (It is available on the web page for this paper.) The similarities (s_{ij}) were converted to dissimilarities (d_{ij}) by $d_{ij} = \sqrt{2(1 - s_{ij})}$. Figure 1 displays the 2-dimensional MDS representation, with the MST overlaid. The MDS used is the representation of the classical metric multidimensional scaling. A 2-cluster single linkage solution is displayed: points in the same cluster are connected by solid edges of the MST, and the dashed line represents the edge separating the two clusters. (On the computer screen color is also used.) Here we see that two clusters are separated between the points 13 and 24, which corresponds to the longest edge of the MST. The only difference between the 2-cluster and 3-cluster solution is that point 22 is separated into its own cluster. We can interactively choose the number of clusters sequentially and watch the process of divisive clustering. (The web page has additional plots illustrating the iterative process.)

Conversely, if we start from 24 clusters and sequentially reduce the number of clusters we can see the steps of agglomeration. At the first step (the 24-cluster single linkage solution) points 2 and 7 are connected, which means the distance of these two points is the shortest among all pairwise distances. Points 21 and 24 are connected in the second step (the 23-cluster single linkage solution).

This is more useful than similarly working through the dendrogram because we can see the relative positions of points using the MDS representation. Also using MST superimposed on MDS we can assess the distortion that exists in a two-dimensional representation of dissimilarity matrix. For example, in Figure 1 the points 3 and 11 are close together in the diagram, but MST shows that point 3 is closer to point 1 than to point 11.

Single linkage cluster analysis is directly related to MST since the distance between groups is defined by the smallest distance between any elements of the two groups, and so MST can be used to visualize the single linkage clustering process easily. However, other hierarchical clustering methods, complete, average, centroid, median and Ward's cluster methods, are not directly related to MST, and so MST cannot be used directly to visualize these clustering process interactively. We modify the representation as follows:

- Within groups, the points are connected by the MST.
- Between groups, a line is drawn which connects the closest elements of the two groups.

Note that, MST is re-computed within each cluster, after sequential splits. Using this modified MST, we can visualize the process of clustering interactively for any hierarchical cluster analysis. The web page shows the 4-cluster complete linkage solution where the points are clustered as (1, 2, 3, 5, 6, 7, 8, 9, 10, 11) (13, 14, 15, 16, 17, 18, 19, 20, 21, 23, 24, 25), (12, 22) and (4). Here dashed lines between groups denote the smallest distance between two points connecting groups. In the 5-cluster complete linkage solution the points are clustered as (1, 2, 5, 6, 7, 8, 10), (3, 9, 11), (13, 14, 15, 16, 17, 18, 19, 20, 21, 23, 24, 25), (12, 22) and (4), which means that cluster (1, 2, 3, 5, 6, 7, 8, 9, 10, 11) is divided by two clusters of (1, 2, 5, 6, 7, 8, 10) and (3, 9, 11) at the next step. Using this sequence we can follow the hierarchical clustering process agglomeratively or divisively and get the difference between steps visually.

For the purpose of comparing agglomeration methods we allow two plots to be made simultaneously. Figure 2 shows the 2-cluster solution of single linkage and complete linkage. Here we can see that using single linkage the two groups are divided by points (13, 24), and for complete linkage the two groups are divided by points (11, 18). It is generally well known that the single linkage method builds clusters by linking objects to those most recently added, producing elongated clusters. On the other hand the complete linkage method concentrates on their internal cohesion, producing spherical clusters. Figure 2 clearly shows the difference between single and complete linkage methods. We can clearly see the difference of two methods as we increase the number of clusters to 5. Generally, it is well known that the complete, average and centroid methods lead to spherical clusters exhibiting high internal affinity, and the median methods weights towards the objects most recently added to the cluster. In our program we implemented six hierarchical cluster methods – single, complete, average, centroid, median and Ward's linkage methods, so we can compare these clustering methods interactively. The web page shows the 5-cluster comparisons of (average, Ward) and (centroid, median) respectively. We can see an interesting fact that several points with only one degree comprise the separate clusters in centroid and median methods. (Degree is the number of edges incident with an observation.) Through interactively running the program, we can also see that in the centroid method points (4, 22, 23, 16) are separated from other points sequentially, while in median method points (4, 23, 12, 22) are separated sequentially, and these two methods are very similar in comprising clusters. From comparing plots like this we can see the differences between several clustering analyses visually.

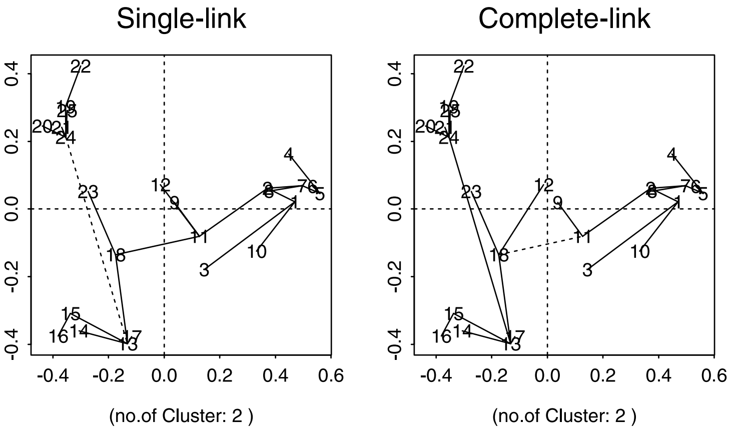

Fig. 2. Single and Complete linkage: 2 clusters

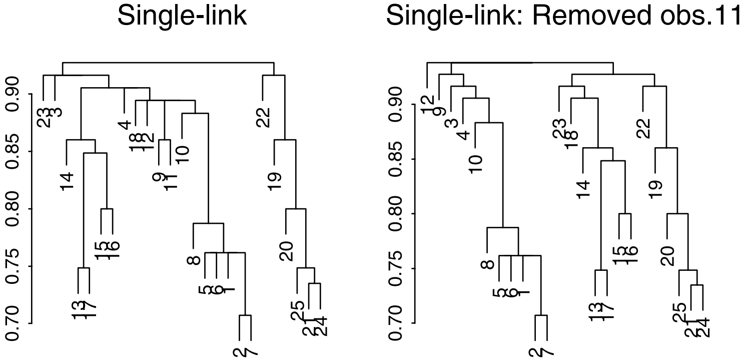

Fig. 3. Dendrogram of single linkage, (left) all 25 villages, (right) without village 11

Detecting influential observations

Cluster analysis is very sensitive to one or two observations, in the sense that their removal may cause large changes in the results of the cluster analysis. Influential observations are defined as those which cause a large change in cluster solution when they are omitted. It is important here to recognize this definition. Points that are influential can be more insidious here than in other types of applications. Outliers to the general point cloud are not necessarily influential, but rather will be peeled off as individual clusters. More influential points can be found in the confluence of clusters, points that fall in "average" positions between cluster centers, or form daisy chains between clusters. Single linkage clustering is especially prone to influence from this type of point. Jolliffe et al. (1995) gives an example showing that removal of a single observation has a substantial effect on the results, using the similarity matrix for the village data. Table 1 shows the 5-cluster single linkage solutions with full data and without observation 11, and dendrograms with full data and without observation 11 are given in Figure 3 respectively. From Table 1, it is clear that the

Table 1. Single linkage cluster analysis on 25 villages – 5-cluster solution with and without village 11

All villages	Without village 11
$\{1,2,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18\}$	$\{1,2,3,4,5,6,7,8,10\}$
$\{3\}$	$\{9\}$
$\{19,20,21,24,25\}$	$\{19,20,21,22,24,25\}$
$\{22\}$	$\{12\}$
$\{23\}$	$\{13,14,15,16,17,18,23\}$

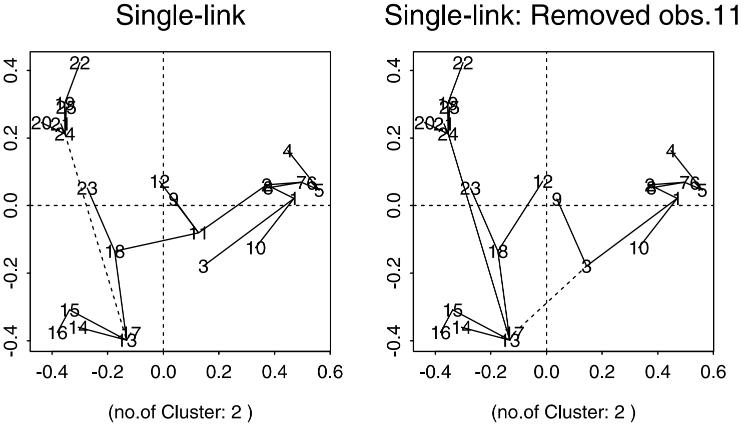

Fig. 4. Single linkage with and without observation 11: 2 clusters

removal of single observation has a substantial impact on the results of single linkage cluster analysis for a 5-cluster solution. Despite large discrepancies near the top of these dendrograms there is a great deal of similarity between two solutions. This is very difficult to see from the dendrograms, and so it is not easy to determine how influential an observation is based on the dendrograms. Jolliffe et al. (1995) considered using the MST to find potentially influential observations in a single linkage cluster. Points with a large degree within suitable radius may have a great effect on cluster analysis.

The approach that we described in Section 2 also helps to visualize influential observations interactively. Figure 4 shows the single linkage solution with and without observation 11 for the 2-cluster solution. For all the data, the two clusters are divided between points $(13, 24)$ and for the data without observation 11, two clusters are divided between points $(3, 17)$. So the solutions are quite different when the observation is excluded, and it is easy to understand what happens: observation 11 is intermediate between 3 and 13 and acts like a connecting link in a chain.

All of what has been described above can be done interactively. The interactive setting helps uncover and understand the nature of influential points, and also helps illuminate how persistent the impact is through numerous stages of the agglomerative clustering process. Information on running the software interactively is available on the web page.

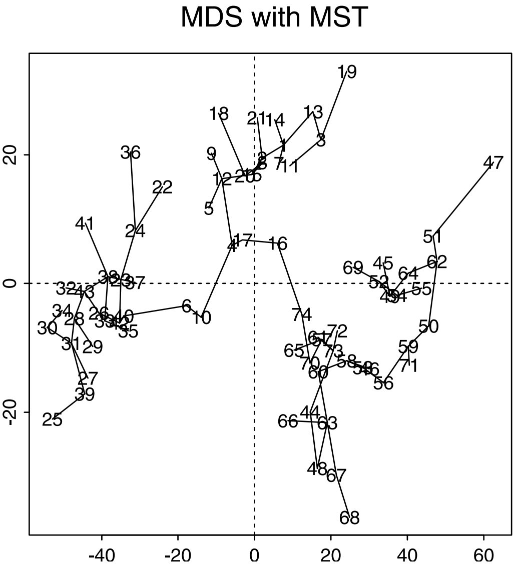

Fig. 5. MDS representation of the flea beetle data: Its not clear to which clusters observations 6 and 10 belong

This approach to assessing sensitivity of the results only can work well for small sample sizes, and primarily for the single linkage method. With complete linkage there is little difference with and without observation 11, which means that observation 11 is not influential in complete linkage cluster. This fact means that whether observations are influential or not depends on the cluster methods. We can also see the effect of observations in other cluster methods.

If we can compare cluster methods in sequential order interactively when observations are removed, it is helpful to see the role of observations in different hierarchical cluster methods. The implemented graphic displays are similar to the previous displays: plots of the solutions on the full data with each linkage method are displayed, alongside the solution without selected cases. (See the web page for examples.) Using this procedure, we can see the effect of observations visually and compare cluster methods after removing some observations interactively.

4 Adding motion

It is important to go beyond 2-dimensional MDS representations to fully understand the inherent cluster structure. To demonstrate this we introduce a second example: flea beetles, first discussed by Lubischew (1962). In this data there are 3 different species of beetles and 6 measured variables, and there are 3 neatly separated clusters. We know the species' identity so it is a good data set to use to test cluster algorithms, especially since the presence of a few influential points confound every hierarchical method. For this data we use Euclidean distance metrics, and results change little if Mahalanobis dis-

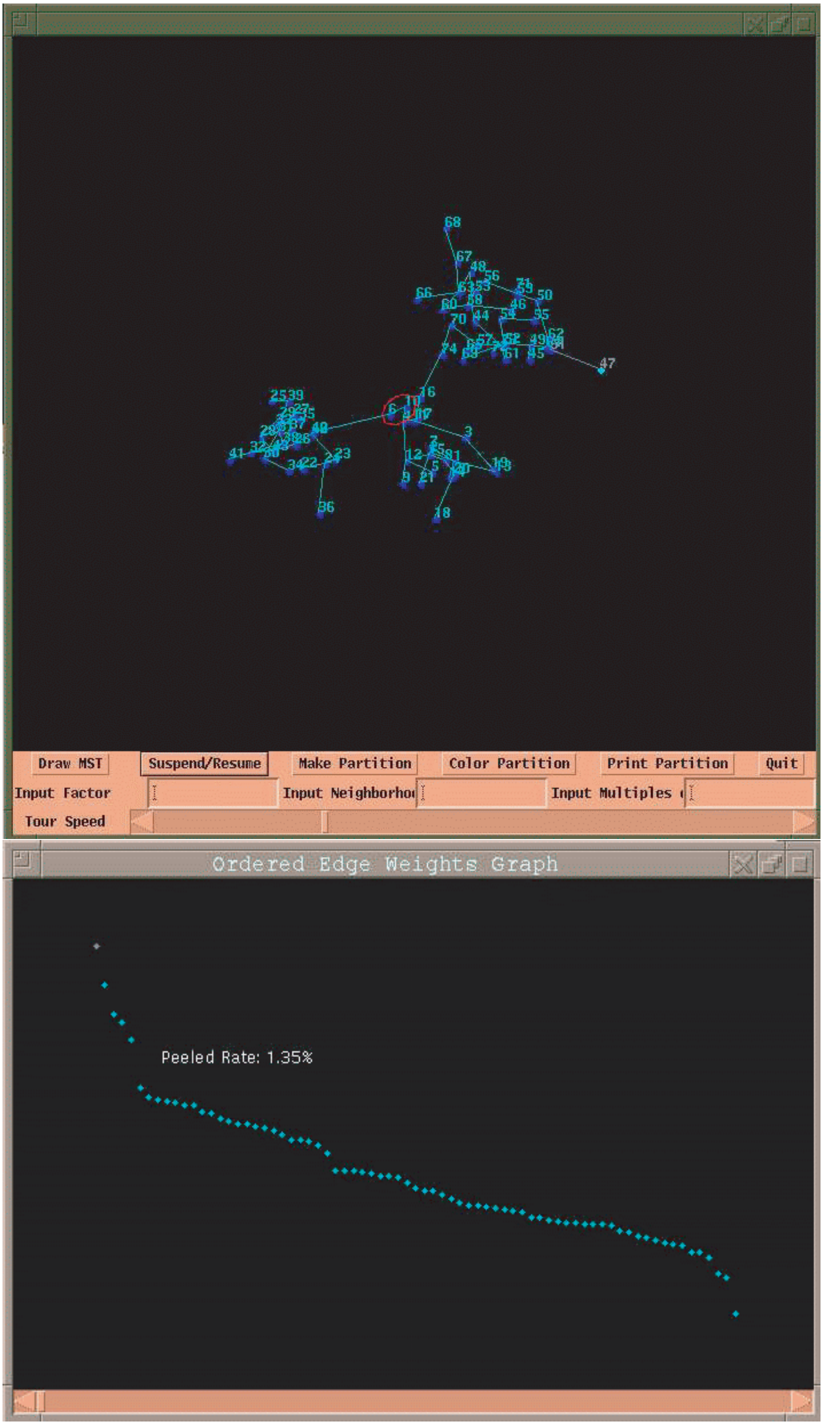

Fig. 6. One projection from the grand tour on the flea beetle data: Cluster identity of observations 6 and 10 more clearly belong to a cluster, the middle one

tance is used. Figure 5 shows the MDS plot, overlaid by the MST. Looking at this plot it is quite easy to recognize the three clusters, although two points (6, 10) are not easily placed in a cluster. But MDS can be misleading because the construction only preserves "relative distance", so points that are "far" in the original space could appear "closer" in the MDS space. Also even with MST overlaid it is not possible to be certain to which cluster the points, 6 and 10, belong.

This confusion can be clarified using a grand tour. In a grand tour the points are placed in motion with a rotation through the original data space. Cluster structure can be identified by similar motion paths or separation of points in some projections. People are surprisingly good at detecting clusters with the grand tour, and can recognize the three clusters in this data set within minutes of viewing. It is a visually simple data set, and yet, there is enough nuisance structure to confound many clustering algorithms. Figure 6 shows one grand tour view where the cluster identity of points 6 and 10 is clear.

We have a prototype JAVA program which runs the grand tour, and also calculates and overlays the MST in the grand tour window. Additionally it displays a graph of the ordered MST edge lengths in the separate window. There is linked brushing between the ordered MST edge length plot and the grand tour window. When the user brushes a point in the edge length plot, the edge is highlighted in the grand tour window. These could be used to "cut" the MST to divide the data into clusters, and in this case the resulting partitions can be saved for further analysis. The partitions made by cutting the MST are the same as those obtained in single linkage clustering, so we can examine the process of single linkage clustering with this program. Our JAVA software can also accept the results of MDS result as input. Hence we can extend 2-dimensional MDS representation to higher dimension, and examine situations when there is only a distance matrix by passing a d-dimensional MDS representation into the grand tour.

In Figure 6 the largest edge length is brushed and it corresponds to the edge connecting points 47 and 51. Trimming this edge would result in a 2-cluster single linkage solution. The second longest edge is between points 6 and 40. This step provides the first "real" split of two clusters, rather than a peeling of outliers. Interestingly, the next real split doesn't come until the 10th longest edge is trimmed, i.e. 11 clusters are made.

Motion can also be used to detect potentially influential points. After noting the major clusters, we can watch for points that fall on the edges of these clusters, especially ones that fall close to other clusters in some projections. For example, points 6, 10, 16, 74, in the flea data. (Note that, points 22, 41, 47 are on the edge of clusters, but on the outer edges and these get peeled off by the hierarchical methods into individual clusters.) To assess impact of the potentially influential points, we remove these points and recompute the clusters. Interestingly, removing observation 6 is sufficient to cleanly split two clusters, and removing observation 74 is sufficient to cleanly split the remaining two clusters (Figure 7). This suggests that removing these two points would enable the single linkage method to perfectly cluster this data.

Expanding on this line of thought, if we were examining complete linkage clustering, from the grand tour we would learn that the clusters are non-spherically shaped and rather elliptical. A natural solution would be to transform the data into principal components, but this doesn't fix the shape problem sufficiently to enable complete linkage clustering to work. The only

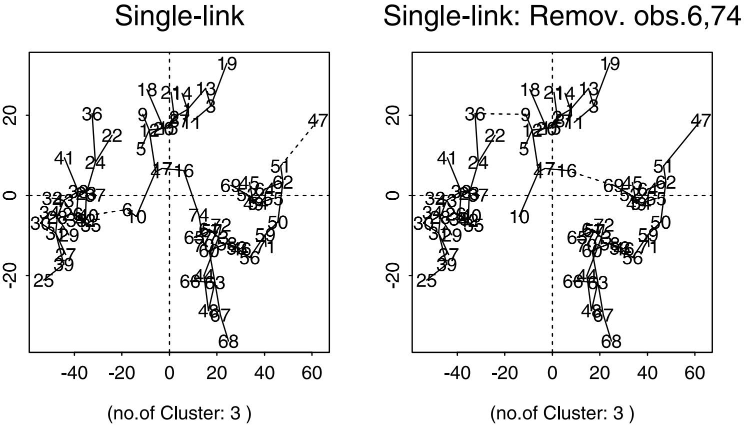

Fig. 7. Three clusters (species) of flea beetle data correctly identified when observations 6 and 74 are removed

approach that facilitates the simple hierarchical cluster methods on this data is projection pursuit, where the data is projected into a 2D solution provided by the Holes index (Cook, Buja, Cabrera, 1993) before beginning cluster analysis. Interestingly, the more complex model-based clustering (Banfield and Raftery, 1993; Fraley, 1999) works perfectly, when different variance-covariance matrices are assumed.

5 Conclusions

Hierarchical cluster analysis can be summarized in a dendrogram, which gives the agglomerative and divisive process. However, it does not provide exploratory representations of data, and it becomes visually unwieldy for even moderate sample sizes. So, we need alternative methods to efficiently compare clustering methods and to see the effect of influential observations in cluster analysis.

In this paper we presented an approach for interactive visualization of hierarchical clusters using MDS and MST, from which we can obtain several benefits related to cluster analysis: (1) we can see the sequential processing of agglomerative or divisive hierarchical clustering interactively, (2) we can compare several cluster methods visually and interactively, (3) we can see the effect of influential observations in cluster analysis interactively, (4) we can examine relationships existing between MDS and cluster analysis visually and (5) we can assess the distortion that exists in a low-dimensional representation of high dimensional data. We also described the benefits of using motion to elucidate cluster structure, and explore potentially influential cases. Having an interactive and dynamic environment can greatly enhance cluster analysis. We hope that clustering software developers might be inspired to incorporate some of these approaches.

It is also possible that the hierarchical methods, especially the single linkage cluster algorithm, through the MST, may be a promising data reduction techniques. We typically consider the k-means algorithm to be the most promising method for reducing the number of cases in very large data sets, however when the data is sparse in high-dimensions partitioning the data using the MST may provide a better reduction (Kwon, 1999).

The implemented S-plus and JAVA source programs, and associated documentation can be obtained from Web site:

www.public.iastate.edu/~dicook/papers/Metrika/paper.html.

Acknowledgments: The authors appreciate the careful reading of the manuscript of three anonymous referees, and the software testing by one referee.

References

Afifi AA, Clark V (1990) Computer-aided multivariate analysis, 2-ed, Van Nostrand Reinhold Co. NewYork

Asimov D (1985) The grand tour: a tool for viewing multidimensional data, SIAM Journal on Scientific and Statistical Computing, 6(1):128–143

Banfield JD, Raftery A (1993) Model-based Gaussian and non-Gaussian clustering, Biometrics, 49:803–821

Bienfait B, Gasteiger J (1997) Checking the projection display of multivariate data with colored graphs, Journal of Molecular Graphics and Modelling, 15:203–215 and 254–258

Buja A, Cook D, Swayne DF (1996) Interactive high-dimensional data analysis, Journal of Computational and Graphical Statistics, 5(1):78–99

Buja A, Swayne DF, Littman ML (1998) XGVis: interactive data visualization with multidimensional scaling, Journal of Computational and Graphical Statistics, To appear

Cheng R, Milligan GW (1996) Measuring the influence of individual data points in a cluster analysis, Journal of Classification, 13:315–335

Cook D, Buja A, Cabrera J (1993) Projection pursuit indexes based on orthonormal function expansions, Journal of Computational and Graphical Statistics, 2(3):225–250

Gordon AD (1981) Classification, Chapman and Hall, London

Fraley C (1999) Algorithm for model-based gaussian hierarchical clustering SIAM Journal on Scientific Computing, 20(1):270–281

Friedman JH, Tukey JW (1974) A projection pursuit algorithm for exploratory data analysis, IEEE Transactions on Computing C, 23:881–889

Gower JC, Ross GJS (1969) Minimum spanning trees and single linkage cluster analysis, Applied Statistics, 18:54–64

Hartigan JA (1975) Clustering algorithms, John Wiley & Sons, INC

Johnson RA, Wichern DW (1992) Applied multivariate statistical analysis, 3-ed, Prentice-Hall, Inc

Jolliffe IT, Jones B, Morgan BJT (1995) Identifying influential observations in hierarchical cluster analysis, Journal of Applied Statistics, Vol 22, No. 1, pp. 61–80

Kruskal JB, Wish M (1978) Multidimensional scaling, Sage Pub

Krzanowski WJ (1988) Principles of multivariate analysis, Oxford Science Pub

Kwon S (1999) Clustering in multivariate data: visualization, Case and Variable reduction. PhD Thesis, Department of Statistics, Iowa State University

Kwon S, Cook D (1998) Using a grand tour and minimal spanning tree to detect structure in high-dimensions, Computing Science and Statistics, 30:224–228

Lubischew AA (1962) On the use of discriminant functions in taxonomy Biometrics, 18:455–477

Morgan BJT, Shaw D (1982) Illustrating data on English dialects, Lore and Language, 3:14–29

Ross GJS (1969) Algorithm AS 13, minimum spanning tree, Applied Statistics, 18:103–104

Ross GJS (1969) Algorithm AS 15, single linkage cluster analysis, Applied Statistics, 18:106–110

Seber GAF (1984) Multivariate observations, Wiley, New York

Swayne DF, Buja A, Hubbell N (1991) XGobi meets S: integrating software for data analysis Computing Science and Statistics, 23:430–434

Swayne DF, Cook D, Buja A (1998) XGobi: interactive dynamic graphics in the X window system, Journal of Computational and Graphical Statistics, 7(1):113–130, See also www.research.att.com/areas/stat/xgobi/.

Zahn CT (1971) Graph-theoretical methods for detecting and describing gestalt clusters, IEEE-Transactions on Computers, Vol. C-20, No. 1, 68–86

Zupan J (1982) Clustering of large data sets, Research Studies Press, A Division of John Wiley & Sons, Ltd, New York, NY