ABSTRACT

This paper details a SAS/GRAPH macro library that aids in identifying the number of natural clusters present in the data. Two macros are available to display the results of nonhierarchical cluster solutions: THREED which generates three dimensional representations of the cluster space andTWOD which generates two dimensional scatter plots. The variables used on the axes are selected by the user and these axes are automatically scaled. Optionally, users may select to represent cluster membership or display the dispersions of all points within a cluster. The use of PROC VISUALS, an experimental SAS dynamic hyperdimensional graphics procedure, can be used as an aid in the selection and interpretation of cluster analysis solutions. Graphs produced by PROC VISUALS and the macro library are similar, but the interactive features of PROC VISUALS make this procedure the superior method of cluster analysis interpretation. Since the availability of PROC VISUALS is limited it is hoped that the macro library will provide most SAS users with the ability to graphically interpret their cluster solutions.

INTRODUCTION

Cluster analysis methods have the common goal of assigning objects to initially undefined classes, "clusters," based on the criteria that individuals within a cluster are, in some sense, closer to one another than they are to other objects outside the cluster. A difficulty in using cluster analysis is trying to optimize the definition of a cluster. As Robert Thorndike pointed out thirty five years ago, what constitutes a cluster of data is an unclear concept:

How, then, shall we decide upon the value of k, the number of families or clusters? Is there any meaningful way of defining an appropriate, or natural, or 'optimum' number of clusters? And once k has been determined, how shall we decide upon the boundaries and the centroids of the various clusters? How shall we tell where one should end and the next begin? Who belongs in a family?

In the years since Thorndike's address no satisfactory definition of a cluster of data has been proposed. Currently it is generally agreed that a cluster of data should possess the characteristics of internal cohesion and external isolation (Cormack, 1971), but no means to optimize this definition is available. The last decade has seen a sharp increase in the use of graphical techniques within statistics in general, and in multivariate analysis in particular (see, for example, Cox 1978, and Everitt 1978, Young et al., 1986). Yet, the techniques of graphical analysis have not extended to the problem of determining the number of natural clusters present in the data. Such a union seems natural since both Thorndike's questions and Cormack's definition depend upon an examination of the clusters and the n-dimensional space surrounding them.

This paper describes two SAS macros written to graphically display data clusters and their surrounding space. By displaying the properties of Cormack's (1971) definition it is hoped that these macros will aid in determining the number of natural clusters present in the data. The macros provide the user the option of selecting which set of variables should be displayed, whether lines should be drawn from each cluster centroid to the data points in that cluster, and whether a circle representing the dispersion of the points should be drawn. This paper also demonstrates how PROC VISUALS, an experimental SAS dynamic hyperdimensional graphics procedure, can be used as an aid in the selection and interpretation of cluster analysis solutions. Graphs produced by PROC VISUALS and the macro library are similar, but the interactive features of PROC VISUALS make this procedure the superior method of cluster analysis interpretation. Since the availability of PROC VISUALS is limited it is hoped that the macro library will provide most SAS users with the ability to graphically interpret their cluster solutions.

DEMONSTRATION DATA SET

The examples used in this paper were generated using PROC FASTCLUS on Fishers' (1936) Iris data. The data is composed of 150 iris specimens each measured on four morphological variables: sepal length and width and petal length and width. The species I. setosa, I. versicolor, and I. virginica were each represented by 50 plants.

THE MACRO LIBRARY

The macro library currently contains the following macros:

1) THREED - displays the clusters as a 2 dimensional scatter plot
2) THREED - displays the clusters in a 3 dimensional representation

These macros may only be used on clusters produced by PROC FASTCLUS. This limitation is the result of the algorithm used by the procedure. PROC FASTCLUS is a k-means algorithm that uses the nearest centroid sorting of Anderberg (1973). Cluster assignment is accomplished by computing the Euclidean distance from each variable to all cluster centroids (SAS, 1988). This assignment process justifies the interpretation of the number of clusters present based on the distance of the data from the cluster centroids.

Before invoking these macros it is necessary to merge the OUT and MEAN data set created by PROC FASTCLUS. These data sets should be merged by cluster number, and variable names in the MEAN data set should be renamed to represent the fact that they now contain cluster centroids and
not raw data.

Once the new data set has been created, the macros can be invoked with a SAS macro call:

\[
\text{XTWOD}(\text{parameter list}); \quad \text{or} \quad \text{XTHREED}(\text{parameter list});
\]

The valid parameters are macro specific and will be explained below. These macros produce no output in either the log or the list files.

**SPECIFICATIONS**

**Macro XTWOD**

The XTWOD macro is designed for preliminary data exploration. The user may graph the cluster solution against any of the variables in the data set. Optionally, group membership of each data point may be identified by drawing lines from the cluster centroid to all points in that cluster. To estimate the boundaries of each cluster a circle around the centroid may be drawn representing the dispersion of the cluster.

PROC GPLOT is the SAS/GRAPH procedure used to plot the data points. If either the lines or the circle option has been requested, a SAS ANNOTATE data set is automatically created to carry out these requests.

The MACRO may be run by placing the following statement in the SAS JOB:

\[
\text{XTWOD(SASdataset,MEANVX,MEANVV, RAWVX,RAWVV,CIRCLE,LINES)};
\]

The options are as follows:

- **SASDATASET** - the SAS dataset created by merging the OUT and MEAN data sets.
- **MEANVX** - the name of the variable containing the cluster centroid of the variable to be plotted on the X axis.
- **MEANVV** - the name of the variable containing the cluster centroid of the variable to be plotted on the Y axis.
- **RAWVX** - the name of the variable containing the raw data values of the variable being plotted on the X axis.
- **RAWVV** - the name of the variable containing the raw data values of the variable being plotted on the Y axis.
- **CIRCLE** - Either 'YES' or 'NO'. Used to determine if circles representing the dispersion of the points around the cluster should be drawn.
- **LINES** - Either 'YES' or 'NO'. Used to determine if lines should be drawn from each cluster centroid to all data points in that cluster.

If the circle option is requested, the radius of the circle is a function of the variances of the data points. Specifically, the variance of the X and Y variables from their respective means is computed and the pooled variance is obtained. In order to project this pooled variance onto the graph, the square root of the pooled variance is calculated. This number is then used as the radius of the circles.

Please note that while these circles do provide a measure of the cluster's internal cohesion, the size of these circles can be decreased by specifying another solution that allows a larger number of clusters to be formed. When determining the number of natural clusters present more attention should be paid to the isolation of the clusters rather than to their dispersion. The internal cohesion measure should only be considered in deciding between solutions that appear to be equally isolated.

**EXAMPLES**

Figure 1 shows the four dimensional cluster solution of the Fisher Iris data. In this graph both the lines and circle options have been selected. In Figure 1, the important component to notice is the lack of isolation in three of the clusters. Of these tightly grouped clusters, the middle cluster appears to be an artificially generated grouping. Not only is this cluster not isolated, but data points from the other clusters overlap. This suggests that there is not a natural division of the data at the displayed boundaries. The cluster may be an artifact of the user requiring a four clusters solution. If a three cluster solution is specified then the data points assigned to this 'artificial' cluster will be divided among other clusters to form more natural boundaries.

Figure 2 shows the three dimensional solution of this data. While the isolation of these clusters is not as extreme as one would like, three clearly separated clusters can be distinguished. The relatively large dispersion of the points around the two upper clusters indicates that these groupings are not cohesive and that these clusters may contain a mixture of the Iris species. The cluster in the bottom left corner is both cohesive and isolated. This suggests that this cluster probably contains a single Iris species.

**Macro XTHREED**

The XTHREED macro was designed for final data exploration. The XTHREED macro provides the ability to display more complex data structures than XTWOD.

Like the XTWOD macro, XTHREED automatically scales the axes and, if requested, lines may be drawn from the cluster centroids to each data point in the cluster. Due to SAS/GRAPH limitations this macro can not draw circles to represent the dispersions of the points around the cluster centroids.

PROC G3D is the SAS/GRAPH procedure used to plot the actual data points. If the lines option has been requested, a SAS ANNOTATE data set will be created.

The MACRO may then be run by placing the following statement in the SAS JOB:

\[
\text{XTHREED(SASdataset,MEANVX,MEANVV, RAWVX,RAWVV,RAWZZ,LINES)};
\]

The options are as follows:
SASDATSET - the SAS dataset created by merging the OUT and MEAN data sets.

MEANX - the name of the variable containing the cluster centroid of the variable to be plotted on the X axis.

MEANVY - the name of the variable containing the cluster centroid of the variable to be plotted on the Y axis.

MEANVZ - the name of the variable containing the cluster centroid of the variable to be plotted on the Z axis.

RAWX - the name of the variable containing the raw data value of the variable to be plotted on the X axis.

RAWVY - the name of the variable containing the raw data value of the variable to be plotted on the Y axis.

RAWVZ - the name of the variable containing the raw data value of the variable to be plotted on the Z axis.

LINES - Either 'YES' or 'NO'. Used to determine if lines should be drawn from each cluster centroid to all points in that cluster.

EXAMPLES

Figures 3 shows the four dimensional cluster solution of the Fisher Iris data. The lack of isolation of the clusters is again evident. A suggested use for this macro is to confirm the decisions reached using the %TWOD macro. A cluster thought to be artificial should show some degree of overlap on the new axis. Such is the case in Figure 3. It is clear that the two dimensional plot is overlapping on the newly added Z axis. Additionally, this graph indicates the span of the artificial cluster across the other axes. These facts support the conclusion that this cluster is not a natural cluster. A more interpretable solution may be obtained if the user recreates the analysis specifying one less cluster.

Figure 4 displays the three dimensional representation of the solution obtained when 3 clusters were requested. Although a limited amount of cluster overlap is apparent, the clusters do generally exhibit the property of external isolation. The conclusion reached from %TWOD regarding the isolation of the clusters is confirmed in this graph. The fact that the clusters contain to remain isolated on the new axis indicates that these are natural clusters.

PROC VISUALS

PROC VISUALS is a highly interactive graphics system that presents a dynamic, high resolution color picture of any data. The data is displayed as a 3D image and it may be thought of as a 3D 'slice' of the data's nD multivariate data space. Users have the ability to direct a moving picture of the data by making the data move and spin. PROC VISUALS dynamic features not only allow the user to rotate the initial view of the clusters, but the 'camera' can also zoom in to explore any overlap that may present. These movements take place in real time, and thus produce a dynamic picture of the data's structure. [For a detailed description of PROC VISUALS see "PROC VISUALS: Experimental SAS Software for Dynamic Hyperdimensional Graphics" in this publication].

In a printed paper it is nearly impossible to convey the dynamic features of this procedure. In the current implementation of PROC VISUALS there is no graceful way to obtain a hardcopy of the view displayed on the screen. For the purpose of this paper, black and white photographs were used. It should be recognized that this medium can not do justice to high resolution color images that PROC VISUALS usually displays. Equivalently, to obtain the figures displayed in this paper literally dozens of other single picture frames had to be examined and evaluated. The information obtained by this dynamic modelling is also lost in this paper.

Figure 5 presents an overview of the three cluster solution of the Iris data. This view is similar to the one presented in Figure 4. Figure 6 was obtained by using the rotation and translation features of PROC VISUALS. The focus has now been shifted to the tightly grouped clusters on the right. Each cluster's data points are identified by a different icon shape. The isolated cluster has now moved to the left side of the picture, and is now longer of concern. Axes and labels have been added to provide some orientation of the view being presented. As this picture shows there is some overlap but easily defined boundary exists between these clusters.

Finally, in Figure 7, the zoom features of PROC VISUALS has been used to examine the boundaries between the two clusters examined in Figure 6. In order to identify the data that did not clearly fall into either of these clusters the data points on the boundaries have been identified by their observation number. This picture confirms the early observation that the clusters do not overlap. It also shows that these clusters do not exhibit a high degree of internal cohesion.

CONCLUSIONS

This paper has outlined how graphical data analysis can be used to interpret the results of nonhierarchical cluster analysis. Two SAS macros that produce graphs of cluster results and an experimental SAS dynamic graphing procedure were demonstrated using the Fisher Iris data. The results of these examples suggest that graphical methods may be a useful means to determine the number of natural clusters present in a data set. How these methods perform with data that contain larger number of natural clusters still needs to be determined.

REFERENCES


Thorndike, R.L. (1953) "Who Belongs in the Family?", Psychometrika, 18, 267-276


Notes
1. SAS and SAS/GRAPH are registered trademarks of SAS Institute Inc., Cary, NC USA

2. PROC VISUALS is an experimental version 6 SAS procedure. The procedure is not currently available for distribution. PROC VISUALS was written at the University of North Carolina, not at the SAS Institute. There is currently no commitment by the Institute to support or distribute PROC VISUALS, nor if it does that it will conform to the description given here. Contact Forrest Young at the UNC address listed below for further information.

3. All graphs in this paper are displayed in black and white although the macros and PROC VISUALS highlight each cluster in different colors.

4. To obtain copies of the macros please contact the author at the address listed below:
   John Larus
   Dept. of Psychology
   Davie Hall 013-A
   University of North Carolina
   Chapel Hill, NC, 27514
   (919)962-7643
   Copies of the macro can be sent through the BITNET mail facility if an electronic mail address is included.
FIGURE 1
4 CLUSTER SOLUTION OF THE RAW DATA

FIGURE 2
3 CLUSTER SOLUTION OF THE RAW DATA