INTERACTIVE GRAPHIC CLUSTERING USING THE PROMENADE SYSTEM

David J. Hall, Geoffrey H. Ball, and Daniel E. Wolf
Stanford Research Institute

1. Introduction

The PROMENADE system is used to explore data using a number of manipulative, graphic, and numeric programming subsystems that are linked together under a single executive program. Interactive interface between subsystems is as important as interactive interface between computer and user.

Principal features of the system are convenience, flexibility, and on-line interaction. Facilities are available for keeping a record of the analysis, using the line printer, still and movie cameras, and a 35-mm microfilm recorder.

This approach to clustering stresses the use of multivariate plots that allow the user, for example, to "fly through hyperspace," observing his data. In addition, multivariate histograms, scatter plots, link-node plots, waveform plots, and others can be dynamically controlled from the console to provide a unique visual insight into data structure, which cannot be obtained from conventional off-line media.

Numeric algorithms that have been developed at SRI are also integrated into the system. One is for cluster analysis and the other for pattern recognition, and both can be interactively controlled from the console. This paper will not discuss the use of the pattern-recognition algorithm. The clustering method known as ISODATA has been available in the system for at least three years. However, as a research topic itself, apart from the existence of PROMENADE as a separate research topic, it is in a fluid state of continual change--hopefully, effecting improvements. The recent additions to the clustering system have been to allow it to be controlled directly from two graphic subsystems in addition to the usual mode of control through its own executive. These three modes of control are all interactive, with the two graphic modes being the most interactive and allowing control of extra details. When the title for this paper was selected, we had hoped to have, at the final writing, some extensive experience at controlling ISODATA from the graphics. However, there were delays in mapping our software system to a new machine configuration, and this paper will therefore be less detailed in this respect. It is also necessary to explain other related features of PROMENADE, before describing interactive control from the graphic subsystems.

Manipulative control of the data is afforded by another subsystem responsible for reading data, in almost any format, and storing it on the disk. This subsystem also provides for selection of a subset of variables or patterns, for performing elementary transformations, and for formation of derived variables, in a convenient way using the keyboard.

The software system structure, using disk overlays, allows the indefinite expansion of the system to eventually incorporate a wider range of perhaps more conventional subsystems and associated program interfaces.

It is predicted that statistical data analysts will prefer on-line interactive graphic analysis in the future. The reasons for this preference will be explained in the following sections.

The PROMENADE system is currently being used on four significant applications:

(1) Chemical--classification of spectrometer data

(2) Marketing--clustering of data on companies

(3) Census--price-index evaluation for housing construction data

(4) Cloud clustering of satellite weather photographs.

a. The Status of Computers and User Modes of Operation

Computers seem to be here to stay, and we have to learn to live with them and make the best use of this powerful tool. So the question is not a matter of choosing between employing computers or not employing them in our statistical data analysis work, but of finding the mode of operation with computers that is best suited to our applications, and the programs that can help us solve our own particular problems. As far as the user operating mode is concerned, there are some broad categories:

(1) Batch computing, characterized by the complete specification, before the computer run, of any control quantities required by the program. A user of a batch program is generally allowed no interaction during the running of the program and should be sufficiently confident of the nature of the program that he is running, probably having used it several times before. The batch user probably requires copious printout, only a small fraction of which is actually read. Most of this printout is obtained in case some additional information is needed.

(2) Interactive, or conversational computing, in which the user works at a teletype keyboard. The user can interrupt the process, or feed it new data while it is in progress, and usually operates in a timesharing computer system. This usually implies a low printout rate of approximately ten characters per second.

(3) Interactive computing, but with graphics, in a dedicated computer

system--i.e., not time-shared. This mode of operation generally implies a high dollar cost per hour for use of the system because while the user is on-line, he is using the whole computer system including card reader and punch, tape units, and line printer in a manner that does not allow sharing of this equipment. Our own facility, a CDC 3300, is now operating in this mode. It is planned to operate in a time-sharing mode early in 1970, but meanwhile, the user of the system must pay approximately $100 per hour.

(4) Interactive graphics with time-sharing, which is a fairly rare commodity but much to be desired. A good system will have sufficient graphical power and relatively rapid response as far as the human is concerned, and because the computer system is being shared, lower costs are incurred than for the dedicated system. Such systems frequently operate around a computer center, with the consoles in fairly close proximity.

(5) Remote graphics computing. A console providing graphic capability is connected to a distant computer, perhaps across the country. This is not generally available today; however, we must expect that such facilities will eventually be available, and plan our data-analysis strategies for the future with this type of network facility in mind. Possible a small computer at the same site as the display console would be an ideal arrangement.

b. The Status of Clustering

Clustering seems to be coming of age as a recognized data-analysis procedure. It is being generally accepted as a means of data compression or summarization, because it automatically places cluster centers in regions of high sample densities so as to represent the full data set by means of a smaller number of prototype, or cluster, points. The success of such a procedure, measured by the error of fit, depends upon the data structure itself. Certain types of data set are therefore more suitable for clustering than others. Because of a range of different types of clustering procedures, and because it is a heuristic rather than an exact method, there are many detailed differences in the clustering methods available.

Related techniques such as clumping, numerical taxonomy, and aggregation are other terms that are used to describe approaches that consider some of the same problems as are faced in clustering. That these techniques are generally coming of age is evidenced by the books of Roach [1], Fisher [2], and Grinker et al. [3]. Many papers on these subjects are reported in the literature, some of which were referenced in a paper by Ball and Friedman [4] on clustering, given at the last annual meeting. A course on clustering is being organized by CEIR, and presented by Ball and Solomon. Of course, the Russians are doing it too (experimenting with clustering), as heard in talks recently given by Professor Aizermann.

There are several batch-processing computer algorithms for clustering that are available on various computers. The ones we know of are our own ISODATA, the Singleton-Kautz program, and the Rubin-Friedman program available through IBM. Thus, by a fairly logical process, if one needs to cluster, a computer must be used, and if some interaction with the program is needed and some graphics output is required for interpretation and control, then some PROMENADE-like system is needed.

c. The General Status of the Current PROMENADE System

We are frequently frustrated by the progress of development of our system, which at times seems to be quite slow. However, in another sense, when we look back we can be partially satisfied with some of the achievements.

Progress is hampered by the following factors:

(1) We seem to be at the forefront of the state of the art in diverse fields. Some of these fields are, clustering and statistical computation, interactive graphics, computer software systems, graphic representations, and the broad field of data analysis.

(2) There are significant basic hardware and software systems problems that must be overcome before we can tackle the problems of our own applications-oriented software.

(3) We require a complex interface and program structure to provide convenience for the user, numerous data-processing methods to match numerous types of data, data-manipulation capabilities, ability to treat input as output, and other features that add complexity to system implementation.

(4) We have had a relatively limited budget for this development.

(5) Upgrading of the computer system from a CDC 3100 to a CDC 3300 has caused problems in the mapping over of system software.

2. General Analysis Procedure

In this section we give an overview of the ways in which the system can be used to analyze data and, more specifically, to perform interactive graphic clustering of multivariate data sets. Later sections of the paper will describe in more detail the features available in each of the subsections.

a. The Users

It is highly desirable, when using the system on new application data, to have at least

two persons to conduct the initial explorations. One of these persons should be an expert, or at least knowledgeable, in the use of the system and the features it can provide. The other person should be knowledgeable about the data collected, its manner of collection, and how to relate the results from the system to some useful conclusions in terms of the application problem. This latter function is usually the more difficult to perform. If a third expert can be obtained, we suggest he should have a strong background in applied statistics.

1) Functions of the Expert System User

If only the expert in the use of the system is present, then he is restricted to working with the data in a purposeless manner, without any specific goals in mind. He must then resort to finding features that he considers may be important. For instance, he may note a certain high value of correlation between a particular pair of variables, because this pair might be the only pair that gives a high correlation. He may notice, for instance, that a particular data point appears, both in the plots and the numeric techniques, to be an outlier--i.e., distinctive. He can identify this point and print out its component values. However, his exploration is aimless, and he may be spending time finding out relationships and characteristics that are obvious to anyone who knows about the application. By definition, a useful exploration must discover something that is not anticipated, and usually experts in some application area already know many facts concerning the data relationships.

2) Functions of the Data-Application Expert

If only the expert in the field of the data application is operating the system, he will initially be unfamiliar with the system and will have to learn its use while he is trying to interpret the meaning of the results. Although it is feasible to use the self-teaching features in the system to guide a new user, we do not advise this approach. If it is not possible to have an expert in the use of the system initially to assist the expert in the data's field of application, we strongly advise some preparation on the part of the data expert before he loads the system. This preparation should consist of reading available reports and documentation.

However, we must stress that, from our past experience, much greater progress is made when a knowledgeable system user and an expert in the field of the data can work together at the console at the same time, as illustrated in Figure 1. The system expert should determine how he can use the system to test the hypothesis formulated by the data expert.

b. Typical Analysis Procedure

The operational procedure that will generally be carried out in order to perform interactive clustering will typically involve the following steps. The system must be loaded into the computer memory and suitable disk files must be opened. This operation generally takes a minute or so, and involves loading a disk pack and pressing a series of buttons that causes about 15 cards to be read at the card reader.

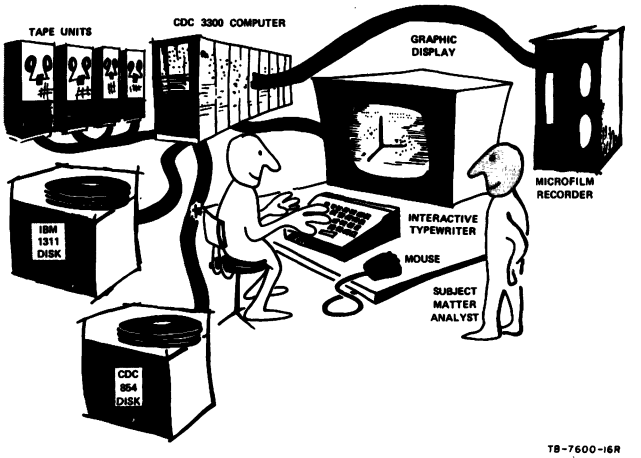


FIGURE 1 TYPICAL MODE OF OPERATION OF THE PROMENADE SYSTEM

The remaining operations can be carried out seated around the graphic console, or going over to the line printer to inspect more detailed output than is available on the display screen.

On the first occasion that a new set of data is read into the system for analysis, a set of data cards must be placed in the card reader and the data on the cards transferred to the disk for more convenient manipulation and semi-permanent storage. Each data file has a name and consists of a two-dimensional data array as shown in Figure 2. The limits on the size of this array

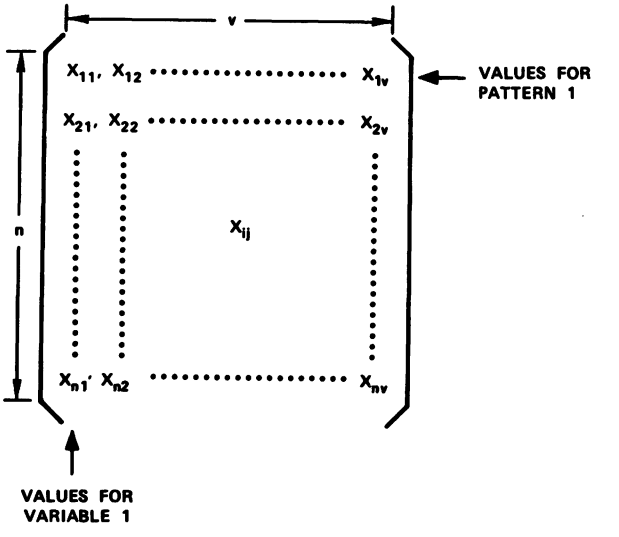


FIGURE 2 DATA ARRAY

are currently fifty variables and one thousand samples. The user of the system chooses which data set he wishes to operate upon by typing in the name of the data set. This data set is then

available for manipulation, creation of derived data sets or subsets, or for listing on the line printer together with some overall statistics. In addition, this data set is automatically linked to the graphic and numeric subsystems.

To cluster the data, the user selects the numeric subsystem, and makes a further selection of the ISODATA clustering routine within that subsystem. He may then proceed to cluster the data, using as initial cluster centers either initial cluster centers that he has punched onto data cards, or internally generated initial cluster centers.

After each iteration of the data set, the clustering program returns control to the user, who may guide the clustering process by changing clustering parameters and performing subsequent iterations. The man/machine dialogue occurs through the media of the graphic screen mainly, or else by inspection of printout on the line printer for fine details. The type of information coming out on the screen may be graphical or numeric, while the line printer is used only for numeric and textual information. In graphic form, there are two types of plots associated with the clustering algorithm by automatic means through the file structure. At the end of each iteration, either the link-node plot or the waveform plot representation of the clustering output can be inspected. Furthermore, the clustering can be controlled by specifications on these plots for further clustering iterations.

In this manner, the man/machine dialogue proceeds until satisfactory clustering is obtained or until the user gives up, or gets too tired. Another more practical termination to the clustering session, but one that unfortunately occurs too frequently, is that it becomes somebody else's turn to use the machine. The bad effect of this annoying interruption is minimized, in the system, by storing intermediate results so that continuation can take place from where one left off, without having to start the procedure from the beginning.

The following sections will give more details of the features available in the various subsystems that are used for interactive graphic clustering.

3. Features of the Data-Manipulation Subsystem

We have found that convenient manipulation of experimental data is a valuable feature of an on-line system. For convenience in handling many files of data, each data set is allocated a 20-character short description and an 8-character file name. Each variable may also have an 8-character mnemonic descriptor that is usually printed or displayed on the CRT screen whenever relevant output appears. The variable number (its sequence) is also used on the printout, which makes it not essential to supply mnemonics. Before these labeling features were implemented, the user frequently wasted valuable time identifying his data entities before he could interpret his output.

The patterns, or samples, have similar numerical identifiers, such as sequence and category labels, etc.

There are 10 flags for each pattern as follows:

(1) Sequence number

(2) Category number

(3) ISODATA Cluster number

(4) Rosen-Hall cluster number

(5) ISODATA flag

(6) Rosen-Hall flag (not used yet)

(7) Character for labeling 3D plot

(8)-(10) Not used.

The following options are available to a system user by typing the corresponding keyboard characters. A list of these options is always displayed to the user if he types "0". A fuller discussion of the general command structure that includes global and local options is given in a more detailed report [5]. The page of options is displayed on the CRT to the user when he types 0. The page format is 64 columns wide by 32 lines high. The user can also have this page printed out on the line printer, if he wishes, and in this case, column and line numbering, as well as page numbering, is also printed out (See Figure 3). This is the chief mechanism for the system designers to update the options pages. This important function in an evolutionary system is done by punching a deck of cards that can be modified and read in again.

```
0000000001111111111222222222233333333334444444444555555555566666666667777777778
1234567891234567891234567891234567891234567891234567891234567891234567890
     DATMAN OPTIONS                                  20 AUG 69          DATMAN   15  1
                                                                        DATMAN   15  2
 C   CREATE A VARIABLE WITH THE ON LINE COMPILER                        DATMAN   15  3
                                                                        DATMAN   15  4
 D   DELETE NAMED FILE FROM THE DISK                                    DATMAN   15  5
                                                                        DATMAN   15  6
 E   GO TO EXECUTIVE                                                   DATMAN   15  7
                                                                        DATMAN   15  8
 F   SELECT A FILE FROM THE FILE DIRECTORY TO PASS TO ALL THE          DATMAN   15  9
       ROUTINES IN THE PROMENADE SYSTEM                                 DATMAN   15 10
                                                                        DATMAN   15 11
 N   READ A NEW DATA SET AND STORE ONTO DISK                           DATMAN   15 12
                                                                        DATMAN   15 13
 P   PRINT THE CURRENT OPENED DATA FILE (PRINTS THE DATA VALUES,       DATMAN   15 14
       MNEMONICS, AND STATISTICS OF A DATA FILE)                        DATMAN   15 15
                                                                        DATMAN   15 16
 S   SELECT A SUBSET OF THE CURRENT OPENED FILE AND STORE THE          DATMAN   15 17
       SUBSET IN A NAMED FILE.                                          DATMAN   15 18
                                                                        DATMAN   15 19
 T   FORM A NEW DATA SET BY TRANSFORMING THE DATA ON THE CURRENT       DATMAN   15 20
       OPENED FILE AND STORE IT ON A NAMED FILE.  TWO CHARS           DATMAN   15 21
       REQUIRED.                                                        DATMAN   15 22
                                                                        DATMAN   15 23
                                                                        DATMAN   15 24
                                                                        DATMAN   15 25
                                                                        DATMAN   15 26
                                                                        DATMAN   15 27
CURRENT OPENED FILE NAME IS                                             DATMAN   15 28
                                                                        DATMAN   15 29
                                                                        DATMAN   15 30
                                                                        DATMAN   15 31
                                                                        DATMAN   15 32
0000000001111111111222222222233333333334444444444555555555566666666667777777778
1234567891234567891234567891234567891234567891234567891234567891234567890
```

TA-5533-9

FIGURE 3 PRINTOUT OF A PAGE OF OPTIONS FOR THE DATA-MANIPULATION SUBSYSTEM

The functions performed by these commands and some additional commands for handling files are next described in more detail. For convenience, the data-manipulation file-handling subsystem is sometimes referred to as "datman."

a. Mode of Operation of Datman Permanent Files

There are 25 permanent files available to the PROMENADE System, with five being occupied by certain data sets used for verifying and exercising the system. These files may not be deleted or overwritten. The remaining 20 are available to System users.

Only one data file will be open and available to all the routines in the System at one time. The file, which is opened by automatic default when first starting the system, is the KENDALL file.

The following commands change the opened file:

F Select a file to use in the System. The file directory is displayed and the operator is required to type in a file name. This name is compared with the names in the file directory. If it matches one of them, the file it matches is opened and the other one is closed. If no match is found, the old open file remains open.

S Select a Subset. The file directory is displayed and the operator is required to type in a new file name. This new file name must be different from the current opened file. If it matches the name of a different file, that file will be overwritten. The operator may then proceed to select a subset of his data set by selecting a subset of the variables. Note that the Command Delete (CD) key on the keyboard deletes the last variable. After selecting variables, the operator selects the indices of a DO loop to select a subset of the patterns.

N Read a new set of data from cards. A deck of data is read from the card reader and stored on the disk. The first card in the data deck contains the file name--i.e., data set name. This card is placed in front of the run-time format card. The file name for the data is compared with the file names in the file directory. If it matches a name in the directory, that file is overwritten with the data. If no match is found, the data set is put on a blank file. If there are no blank files, the operator must delete a file before reading the data.

T Transform the data. This works much the same as the select-a-subset option. It displays the file directory and allows the operator to type in a file name under which to put his transformed data. This name must be different from the name of the currently opened file. The name is compared to the names in the file directory. If it matches the name of a file, that file will be overwritten. The operator then must select the variables to be transformed. This is done by pointing at the variable mnemonics on the screen with the mouse and pressing the button on the mouse to select the variable. Typing Command Delete (CD) deletes the last variable that was selected. Type "space" as soon as the list of variables is complete. The operator then must select a transformation by typing one of the following characters:

A Reciprocal of square root $X_{new} = 1/\sqrt{X_{old}}$

E Exponential $X_{new} = e^{Xold}$

L Logarithmic $X_{new} = Log_e X_{old}$

M Fill in missing data--type in a value. All data found equal to it will be replaced with the average for that variable (the average is calculated with the value and all like it deleted).

N Normalize the data--mean to 0.0 and standard deviation to 1.0

Q Square $X_{new} = (X_{old})^2$

R Reciprocal $X_{new} = 1/Xold$

S Square root $X_{new} = \sqrt{X_{old}}$.

At the end of the transformation the new file containing the transformed data is left open and the file containing the untransformed data is closed.

C Create a new variable. This allows the user to derive new variables from any combination of existing ones, using arithmetic and functional forms as well. After typing C, the user must enter the name of a new variable. This name must be in the format of a standard FORTRAN identifier. Then must follow a relational expression using at least one already defined variable, and any other already defined variables that the user may wish. For example:

```
NEWVAR3 = PETLEN * 2.15
+ (SEPLEN ** 2)/9.866
```

would be a valid expression. This extra variable would be automatically added as an extra column to the data file, after the necessary calculations had been carried out for each row of the data set. The new expression is also printed out on the line printer as a permanent record.

The following command may change the opened file:

D Delete a file. This command allows the operator to delete a file from the disk. The file directory is displayed. The operator must type in the name of the file he wants to delete. The name is compared with the file names in the file directory. If it matches the name of a file in the directory, the file is deleted from the disk. If it matches the name of the currently opened file,

the file is deleted and the KENDALL file is left open. If no match is found, a message is printed and nothing happens in the program.

The following commands cause no change in the opened file:

P Print the data. This command prints the file header, data statistics, and data values of the currently opened data file.

E Exit. This command returns control to the Executive Subsystem of the PROMENADE System.

These data-manipulation features mainly provide convenience to the user. Similar functions can be performed off-line, or by re-punching cards, etc., but only at the cost of a much greater delay. In fact, the difference in delay is so many orders of magnitude that the ability to perform these manipulations on-line and then immediately use the results is a qualitative difference rather than a quantitative one in effect. If the manipulations were not so convenient to do, they probably would not get done by the other, slower means.

4. Features of the Numerical Clustering Algorithms

The clustering algorithms presently integrated into the system are the ISODATA clustering algorithm and the Rosen-Hall pattern-recognition algorithm. This latter performs a clustering, subject to the constraint that all patterns composing a cluster must have the same category label. The ISODATA algorithm, due to recent improvements to the system, can be controlled from three different points as shown in Figure 4.

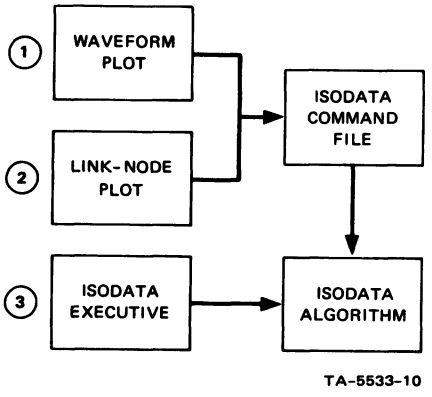


FIGURE 4 THREE CONTROL MODES FOR ISODATA

The controls from the two plots are treated in more detail in the following section. Briefly, they allow the user to watch the progress and control the clustering while viewing the plots, rather than having to go backward and forward from the ISODATA executive to the plots--a process that takes about 10 seconds and is annoying to perform between each iteration.

The ISODATA executive allows for overall control of the algorithm, for parameter changes, printout controls, etc., as fully documented in other reports [5]. The clustering performs grouping of data points that are close together in multivariate space. This data-compression or fitting process is useful when trying to summarize data. Although there are many types of summary statistics and graphs that are produced as output from the clustering, perhaps the most global information is obtained from the clustering characteristic curve, as shown in Figure 5.

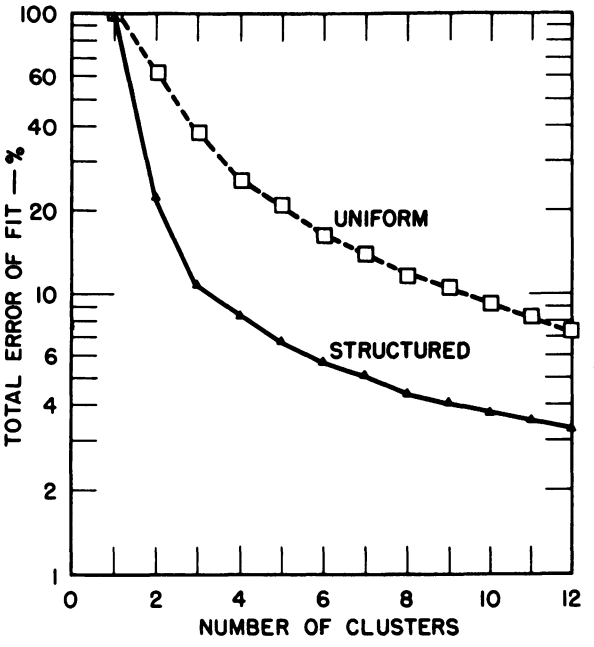
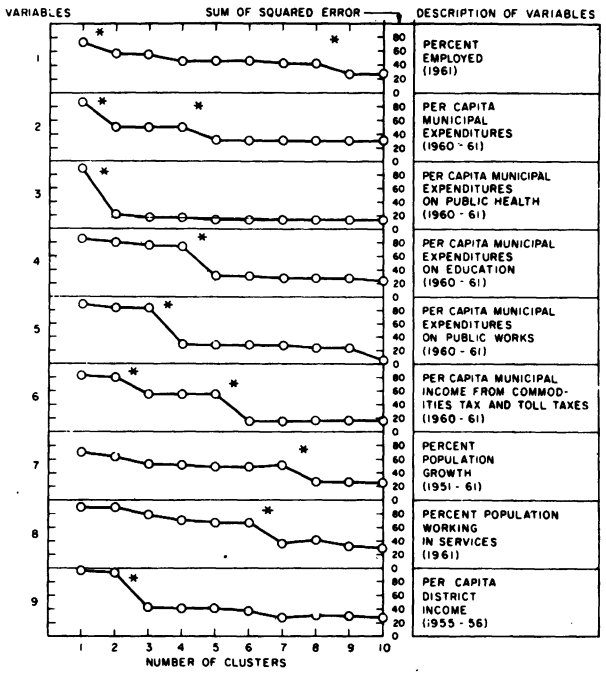


FIGURE 5 CLUSTERING CHARACTERISTICS FOR UNIFORM AND TYPICAL STRUCTURAL DATA

The way in which each individual variable contributes to the total error of fit can also provide useful information, as shown in Figure 6. This is a convenient way of showing which variables take part in the partitioning of clusters, and at what stage of the process. Other information that is useful appears on the display screen in textual and numeric form. For example, the iteration number, number of clusters, parameters and printout values, etc. are displayed. More detailed results appear on the line printer, such as the positions of all the cluster centers, and which patterns they contain, how many are in each cluster, and what the average distance (dispersion) is in each cluster. The user can choose plotting options from the ISODATA executive, either the link-node or waveform plots, to view the results. Linkage of the clustering output data as input to the plots is done automatically by the system. To "go" to one of the plots, the user must type "E", to return to the executive of the system, and then type "L" for link-node (see Figure 7), or "W" for waveform (see Figure 8).



TA-5533-12

FIGURE 6 CLUSTERING CHARACTERISTICS FOR INDIVIDUAL VARIABLES

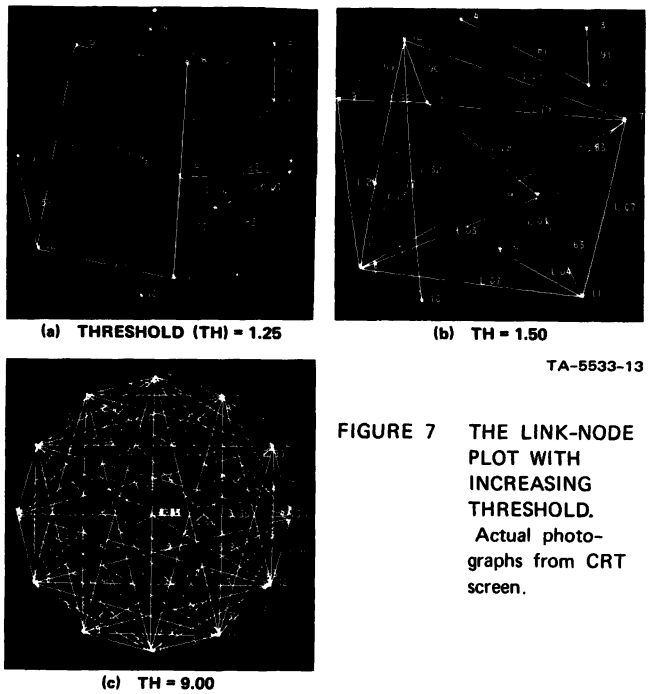
The ISODATA algorithm has been steadily developed over a period of about five years, from a batch program to an on-line interactive version with associated graphics. Both ALGOL and FORTRAN versions for many different machines have been developed. As clustering continues to be useful in many applications we find that we can provide new features to give greater insight and control, and to carry out the clustering more rapidly, with greater conveneience.

The user can review the options available to him by typing O, and in this case, the options as given in Figure 9 will be displayed. The IPRINT parameter is displayed on the parameter control page. If the value of IPRINT is changed from its initial value of zero, fairly extensive printout will be obtained on the printer. If the value of IPRINT is set to 11, then, in addition to the normal printout, the distance of each pattern from its cluster center is also printed. In order to accommodate a wide range of magnitudes of numbers, because this is designed to be a general-purpose system, an internal calculation is done on numbers before they are printed. Depending on the results of this calculation, an appropriate run-time output format is chosen. This avoids the cumbersome form of output sometimes seen used in other systems where 8995.2600000E-3 is used instead of 8.99526.

5. Graphic Control of Clustering

a. Graphic Features in PROMENADE

Graphical representation of multivariate data is achieved in the system through several types of plots. These plots usually provide automatic scaling and labeling features, automatic



TA-5533-13

FIGURE 7 THE LINK-NODE PLOT WITH INCREASING THRESHOLD. Actual photographs from CRT screen.

linking to data files, and can generally plot data faster than a human can absorb the meaning of the plots. This level of automation is a significant advantage over other techniques. The interactive nature of the plots, and the ability to hunt for a good view, allows for selection and capture, in hard-copy form, of only significant information. In the batch approach, a much larger volume of hard copy information usually has to be captured in order to find the significant information. Providing the user with effective graphic insight to multivariate data is quite a challenge to the display programmer. It is also evident that a single type of graphical transformation will not be adequate for general data-analysis purposes. Hence, in PROMENADE, numerous plots are available. Currently, the following plots are provided:

(1) Scatterplot, or two-dimensional plot

(2) Multivariate histogram plot

(3) Distance-along-versus-away-from a line plot

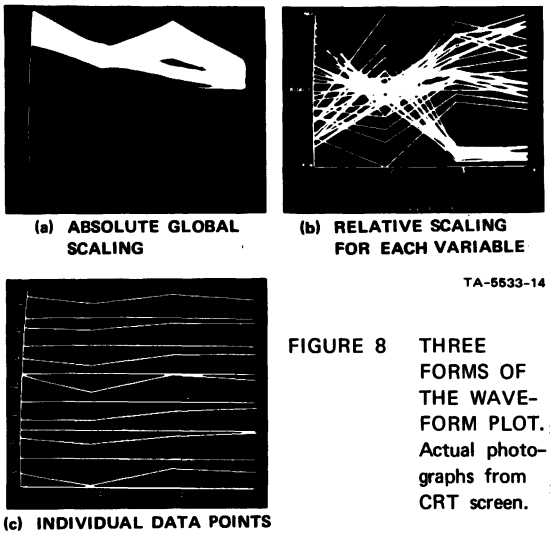
(4) Psuedo three-dimensional plot

(5) Pseudo four-dimensional plot

(6) Waveform plot, or profile plot

(7) Link-node plot

These plots are mostly of a dynamic nature, and each can be controlled by numerous control parameters. For example, the pseudo four-dimensional plot allows the user to get a true perspective view of his data as though he was flying about in it. Although all these plots can aid exploration of a data set that is to be clustered, and thus indirectly aid the purposes of clustering, direct control of the ISODATA clustering has only been implemented via the latter two plots.



(a) ABSOLUTE GLOBAL SCALING

(b) RELATIVE SCALING FOR EACH VARIABLE

TA-5533-14

(c) INDIVIDUAL DATA POINTS

FIGURE 8 THREE FORMS OF THE WAVE-FORM PLOT. Actual photographs from CRT screen.

b. Graphic Control Through Both Plots

Examples of both plots are shown in Figures 7 and 8. Each plot has a set of options or parameters that can be changed to change the plots themselves. In addition they have commands for controlling the ISODATA clustering. These latter commands are identical, but each plot has its own unique plot parameters for controlling the data representation. For example, Figure 10 shows a printout of the page of options for the waveform plot. This page appears on the CRT screen when the user types "O".

```
0000000001111111111222222222233333333344444444445555555556666666667777777778
1234567890123456789012345678901234567890123456789012345678901234567890123456789
AT THE BEGINNING OF EACH ITERATION, YOU MAY CHANGE THE VALUES OF  ISODATA  23  1
THE LAST 7 PARAMETERS SHOWN ON THE PARAMETER DISPLAY.  CHOOSE A   ISODATA  23  2
NUMERIC PARAMETER WITH THE MOUSE AND TYPE IN THE NEW VALUE.       ISODATA  23  3
THE LAST TWO DIGITS TYPED FOR THETAC WILL BE TAKEN AS THE         ISODATA  23  4
FRACTIONAL PART.  OTHER OPTIONS ARE...                            ISODATA  23  5
                                                                  ISODATA  23  6
I       REINTRODUCE INDIVIDUAL PATTERNS INTO THE CLUSTERING SET   ISODATA  23  7
D       DELETE INDIVIDUAL PATTERNS FROM THE CLUSTERING SET.       ISODATA  23  8
X       PUNCH OUT THE CLUSTER CENTERS AND THE PCM TABLE.          ISODATA  23  9
E       RETURN TO THE EXECUTIVE.                                  ISODATA  23 10
O       DISPLAY THE LIST OF OPTIONS.                              ISODATA  23 11
                                                                  ISODATA  23 12
COMPUTATIONS ARE BEGUN BY TYPING ONE OF THE FOLLOWING VALUES FOR  ISODATA  23 13
THE SPLIT/LUMP PARAMETER...                                       ISODATA  23 14
                                                                  ISODATA  23 15
S       SPLIT...A TRIAL SPLITTING OF CLUSTERS IN TWO IS           ISODATA  23 16
            PERFORMED AND THE MEANS OF THE SUB-CLUSTERS           ISODATA  23 17
            COMPUTED.  IF THE MEANS ARE MORE THAN 1.1THETAC       ISODATA  23 18
            UNITS APART, THE NEW CLUSTERS ARE KEPT AND THE OLD    ISODATA  23 19
            ONE DELETED.                                          ISODATA  23 20
L       LUMP...TWO CLUSTERS ARE COMBINED IF THE MEANS ARE LESS    ISODATA  23 21
            THAN OR EQUAL TO THETAC UNITS APART.  AT MOST         ISODATA  23 22
            NCLST PAIRS OF CLUSTERS WILL BE LUMPED IN AN          ISODATA  23 23
            ITERATION.                                            ISODATA  23 24
N       NEITHER SPLIT NOR LUMP.                                   ISODATA  23 25
A       THE PROGRAM WILL DECIDE WHETHER TO SPLIT OR LUMP.         ISODATA  23 26
SPACE   USE THE PREVIOUS VALUE OF THE SPLIT/LUMP PARAMETER.       ISODATA  23 27
                                                                  ISODATA  23 28
CLUSTERS WITH FEWER THAN THETAN PATTERNS ARE DELETED.             ISODATA  23 29
FOR NO LINE PRINTER OUTPUT, SET #IPRINT# TO ZERO.                 ISODATA  23 30
                                                                  ISODATA  23 31
PRESS THE SPACE BAR TO REDISPLAY THE PARAMETERS.                  ISODATA  23 32
0000000001111111111222222222233333333344444444445555555556666666667777777778
1234567890123456789012345678901234567890123456789012345678901234567890123456789
```

TA-5533-15

FIGURE 9 COMMAND OPTIONS FOR ISODATA

A typical command structure for controlling ISODATA from this plot can be written in general form as:

XY (parameter string) (command accept)

where

XY is one of the listed two-letter commands

(parameter string) is a list of cluster or pattern numbers separated by commas

(command accept) is a special keyboard character.

For example, the command:

KS 5,7,14 (command accept)

would cause splitting of the clusters 5, 7, and 14.

c. Interpretation Using the Waveform Plot

The waveform plot is more familiar to psychologists as the profile plot. In Figure 7 examples of this plot in its various forms is given. The figures show a four-dimensional data set. In Figure 8(a) the scaling is global--i.e., the same scale is used for all dimensions and the data is placed symmetrically about a zero-valued center line. The maximum data value is displayed at the maximum height of the vertical scale. The number of samples on the screen/total number of samples in the data set, is also given. Provisions for selecting subsets of data are made in this subsystem. Figure 8(b) is similar except for the scaling, which is local to each dimension. This allows a more detailed examination of the range of data values in each dimension but does not allow magnitude comparisons from one dimension to the other. The scaling is arranged to place the maximum and minimum values at the extremities of the vertical axis. The ability to visually observe clustering of data, and to see which dimensions are contributing to clustering, is the chief advantage of these superimposed waveform plots.

In Figure 8(c) the data is displayed by individual data point, and numbered sequentially on the left to allow identification. Scaling can be either global or local in this plot also. When the user depresses the space bar, the next group of seven data points is shown. Cluster centers can also be displayed in these same formats since they are prototype data points, but they are displayed using dotted lines to distinquish them.

d. Interpretation Using the Link-Node Plot

In the current version of the link-node plot only the cluster centers are displayed. These are initially placed on the circumference of a circle, as shown in Figure 7(c). Initially no links are displayed--only nodes and their corresponding numbers. This is a non-geometric plot in that multivariate data values are arbitrarily placed initially, and in that many dimensions are "squashed" down to two. A threshold value can be introduced to create links. Links are drawn between any two nodes if their distance apart in the original multivariate space is less than the value of threshold used. This plot can also be described as a way to graphically represent a two-way table of distances. Since the CRT screen is a volatile dynamic medium, it is easier to perform these operations in this way rather than by pencil and paper.

In Figure 7(a), the initial positions of the nodes have been moved, and the threshold value (TH) is set to 1.25 by the operator. Note that there are two connected networks of points, and one isolated cluster center, number 10, at

this threshold value. Raising the threshold value to 1.50, as shown in Figure 7(b), causes more links to be drawn. This figure also illustrates some rearrangement of node positions by the operator. Now cluster center 10 is linked, but the two networks are still separated. In Figure 7(c), all possible pairs of links are drawn with a threshold value of 9.00.

6. Applications

Applications can be divided into two groups:

(1) A small application, involving typically a few days of discussion and interpretation, and a few hours of computer time.

(2) Significant applications, involving long-term contractual agreements and substantial amounts of effort and money.

In Category 1, we have processed the following data:

- Agricultural data for pineapple growth
- Economic factors for Indian cities
- Nutritional data for Hawaiian school children
- Socio-economic factors for gasoline station placement, etc.

In Category 2, four applications are in progress:

- Chemical, for classification and clustering of spectrometer data
- Marketing, to classify companies from their public financial data
- Census, comparison of regression and clustering methods for housing-construction price index
- Cloud clustering from satellite photographs.

It is not the purpose of this paper to discuss applications of interactive clustering in detail. A paper will be presented discussing the four applications in Cateogry 2 in detail [6].

7. Conclusions

We have a significant facility for performing interactive graphic clustering. The last four months have been dedicated to changing over to the new computer. We hope the next phase will be one of active use of the system on significant applications. Meeting the needs of the application problems is the ultimate function of the system, and many ideas for system improvement are generally gained in trying to better meet the application needs.

```
0000000001111111111222222222233333333334444444444555555555566666666677777777778
1234567890123456789012345678901234567890123456789012345678901234567890123456789012
     OPTIONS FOR WAVEFORM PLOT                          20 AUG 69  WAVEFORM 33  1
                                                                   WAVEFORM 33  2
                                                                   WAVEFORM 33  3
     G    PERMUTE VARIABLES                                        WAVEFORM 33  4
     P    SELECT POINTS FOR THE PLOT                               WAVEFORM 33  5
     O    DISPLAY THIS PAGE OF OPTIONS                             WAVEFORM 33  6
     M    CHANGE MODE OF PLOT                                      WAVEFORM 33  7
     E    EXIT                                                     WAVEFORM 33  8
     S    CHANGE SCALE FOR PLOTS                                   WAVEFORM 33  9
                                                                   WAVEFORM 33 10
                                                                   WAVEFORM 33 11
                                                                   WAVEFORM 33 12
                                                                   WAVEFORM 33 13
                                                                   WAVEFORM 33 14
          THE FOLLOWING COMMANDS CONTROL THE ISODATA CLUSTER       WAVEFORM 33 15
                    ANALYSIS ALGORITHM.                            WAVEFORM 33 16
                                                                   WAVEFORM 33 17
     KB   BEGIN ISODATA CLUSTERING -- TYPE IN SPHERE FACTOR       WAVEFORM 33 18
     KC   CREATE A CLUSTER CENTER                              NI WAVEFORM 33 19
     KD   DELETE THE SPECIFIED CLUSTER(S)                         WAVEFORM 33 20
     KF   FREEZE THE SPECIFIED CLUSTER(S)                         WAVEFORM 33 21
     KG   GO TO ISODATA AND DO SOME SETTLING ITERATIONS           WAVEFORM 33 22
     KL   LUMP THE SPECIFIED CLUSTERS                            WAVEFORM 33 23
     KS   SPLIT THE SPECIFIED CLUSTER(S)                         WAVEFORM 33 24
     KU   USE THE SPECIFIED CLUSTER(S)                           WAVEFORM 33 25
     PD   DELETE THE SPECIFIED PATTERN(S)                    NI WAVEFORM 33 26
     PF   FREEZE THE SPECIFIED PATTERN(S)                    NI WAVEFORM 33 27
     PU   USE THE SPECIFIED PATTERN(S)                       NI WAVEFORM 33 28
     VI   IGNORE THE SPECIFIED VARIABLE(S)                   NI WAVEFORM 33 29
     VU   USE THE SPECIFIED VARIABLE(S)                      NI WAVEFORM 33 30
                                                                   WAVEFORM 33 31
                                        NI = NOT IMPLEMENTED       WAVEFORM 33 32
0000000001111111111222222222233333333334444444444555555555566666666677777777778
1234567890123456789012345678901234567890123456789012345678901234567890123456789012
```

FIGURE 10 OPTIONS FOR WAVEFORM PLOT

References

1. Roach, S. A., "The Theory of Random Clumping," Methuen and Co., London (1968), pp. 94.
2. Fisher, W. D., Clustering and Aggregation in Economics, Johns Hopkins Press, Baltimore, Maryland (1969).
3. Grinker et al., The Borderline Syndrome, Basic books (1968). See especially Chapter 5.
4. Ball, G. H. and Friedman, H. P., "On the Status of Applications of Clustering Techniques to Behavioral Sciences Data" Proceedings of the Social Stats. Sec. (1968), pp. 34-39.
5. Hall, D. J. et al., "Promenade--An Improved Interactive-Graphics Man/Machine System for Pattern Recognition," Final Technical Report, RADC-TR-68-572, Rome Air Development Center (June 1969).
6. Hall, D. J. et al., "Applications of the PROMENADE Data-Analysis System," submitted to Computers and Communications Conference, Mohawk Valley Section IEEE, Rome N. Y., 1 October 1969.