Recent Developments in Clustering and Data Analysis

Développements Récents en Classification Automatique et Analyse des Données

Proceedings of the Japanese-French
Scientific Seminar
March 24–26, 1987

Edited by

Chikio Hayashi
University of the Air
Wakaba, Chiba
Japan

Michel Jambu
Centre National d'Etudes des Télécommunications
Issy Les Moulineaux
France

Edwin Diday
INRIA
Domaine de Voluceau
Le Chesnay Cedex
France

Noboru Ohsumi
The Institute of Statistical Mathematics
Minato-ku, Tokyo
Japan

ACADEMIC PRESS, INC.
Harcourt Brace Jovanovich, Publishers
Boston San Diego New York
Berkeley London Sydney
Tokyo Toronto

ROLE OF COMPUTER GRAPHICS IN INTERPRETATION OF CLUSTERING RESULTS

Noboru Ohsumi

The Institute of Statistical Mathematics
4-6-7, Minami-Azabu, Minato-ku, Tokyo, JAPAN

I. INTRODUCTION

The utility of computer graphics has now won the recognition of many researchers specializing in data analysis, and a variety of graphical representation methods have been proposed by them. What data analysts expect most from graphical representation is to obtain a firm grasp of the intrinsic structure of collected data by visualizing their features.

The hardware environment bears very closely upon free use of graphical representation as a means of data analysis. In this connection, it deserves attention that the functional improvement of microcomputers and workstations, realized recently in upgraded graphics functions as well as increased performance in peripheral equipment, has provided a wider freedom of graphical representation and greater ease of graphics equipment operation.

These graphical representation methods have some common features: (1) representing the features of multivariate data, (2) visual inspection of the relationship among variables, (3) observing data distribution, (4) projections based on data transformation and features selection, and (5) enhancement of intrinsic features in data.

ISBN 0-12-215485-1

II. GRAPHICAL REPRESENTATION IN AUTOMATIC CLASSIFICATION

Graphical representation is also applied for visualizing the results of automatic classification. While these can be represented on a computer graphics display with little difficulty, the interpretation of displayed graphs and figures is not as easy as the observation of histograms and scatter plots and tends to be influenced by subjective judgement. It is for this reason that computer graphics leaves something to be desired despite the increasingly important role it is playing in automatic classification.

In most cases, these representations of classification results must be considered the preliminary stage to developing more suitable graphical representation methods for grasping the meanings of classification. At the same time, many of clustering methods are heuristic in nature, and importance is attached to the comparison between classification methods as well as the comparison of classification results. Thus, what the analyst expects from classification can be summarized as follows.

1. Immediate visual inspection of classification results.
2. Fast recognition of what is being classified in what way.
3. Easy identification of the class to which an individual belongs.
4. Simplified recognition of the variables contributing to classification in multivariate data.
5. Rapid understanding of the differences or similarities between clusters by visual inspection.

However, in fulfilling these requirements with computer graphical methods, a broad overview and consistent approach is imperative. The assumptions, algorithms, and procedures

underlying a particular computerized representation must be carefully examined to assure compatibility and consistency with other graphical methods used by the analyst in his analysis. Failure to maintain such consistency in both individual modules and, especially, in integrated packages will tend to magnify the subjective element inherent in visual interpretation of graphical representations. Hardware consistency is also required. The hardware environment must be constructed systematically to optimize the capabilities of the software system.

In addition to consistency, the ideal system should have flexibility, allowing the analyst to select from a wide variety of graphical representation methods and conventional multivariate methods and to do so rapidly and easily.

For the past several years, we have been working on the development of such an integrated, consistent hardware-software system for use in performing automatic classification and in visualizing and interpreting the results.

III. CONCEPT OF COMPUTER GRAPHICS SYSTEM IN AUTOMATIC CLASSIFICATION

Because of the phenomenal development in computer technologies in recent years, it is now possible to meet all standard requirements for data analysis using terminals connected to large-scale computers or microcomputers. However, in order to meet the demands of data analysts as described above, development of a computer graphics system designed specifically for the use of experts is much desired. A project started some years ago provided an opportunity for us to embark on the development of such a computer system in 1982. We started with a minicomputer because the powerful

microcomputers seen today were not available at that time. The system has since been refined and upgraded by stages and is now a powerful system using graphics. It has been given the name Automatic Classification Techniques for Interpretation, Visualization and Evaluation, or simply, the ACTIVE workstation.

A. System Design Policy and Features

The ACTIVE workstation has been built in stages. At first, a minicomputer with intelligent color graphics display was introduced as an intelligent terminal connected to a large-scale host system. Next, the peripheral equipment were strengthened by introducing color XY-plotters, tablet digitizers, color video cameras, video recorders, image recorders, etc. Recently, professional workstations, graphics server processor and high-resolution color display were added, and as completed, the system consists mainly of the workstation. Figure 1 shows the configuration of the ACTIVE workstation. The ACTIVE workstation has the following features.

1. The workstations and minicomputer can be used independently in the system.
2. Both serve as intelligent terminals to the host system.
3. Each of the special microprocessors built into the color graphics displays is available for image analysis and provides intelligent functions.
4. A channel controller links the various components to each other and to a host system, allowing free access to any part of the system.
5. Files can be freely transferred between subsystems.
6. Additional microcomputers can be linked to the workstation

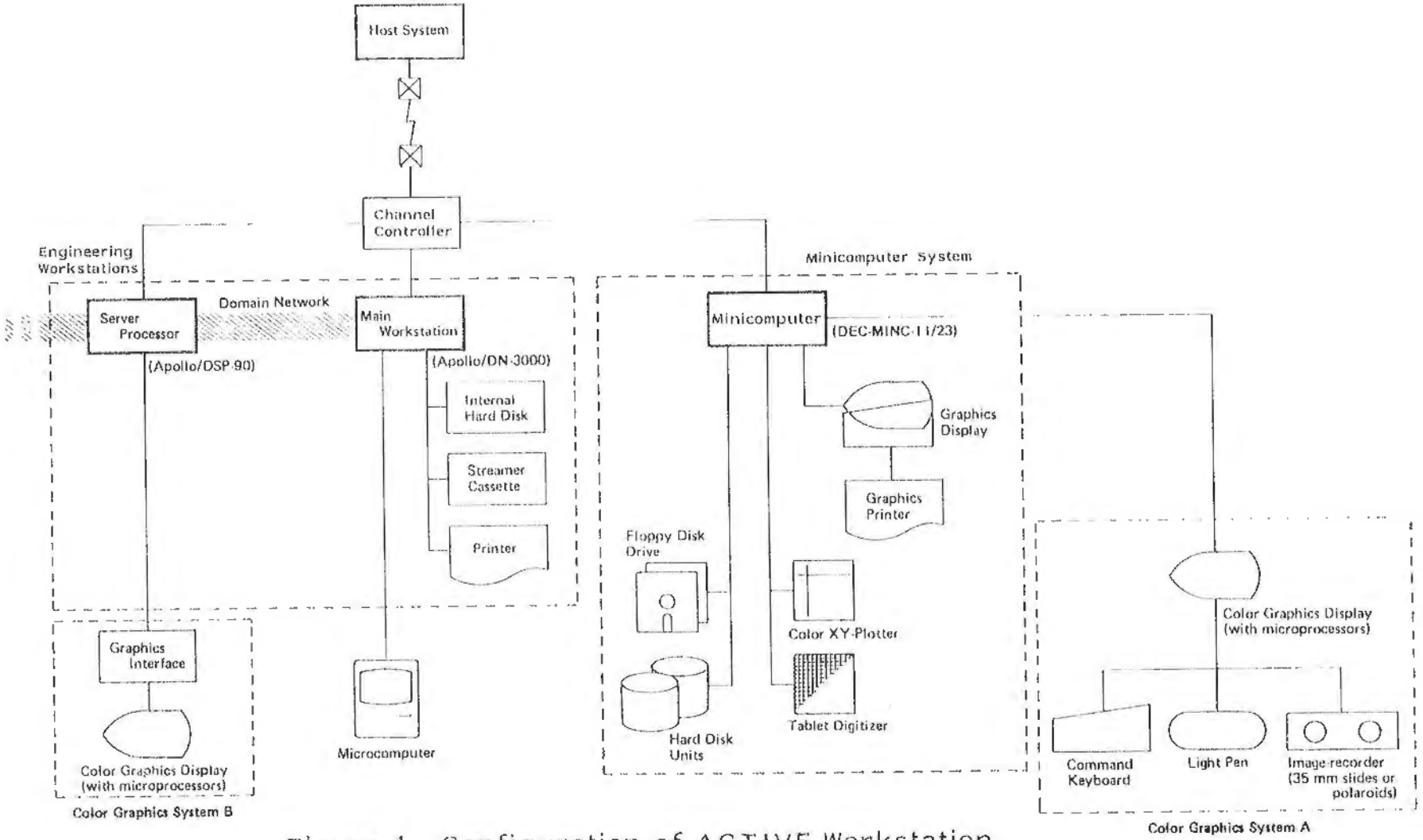


Figure 1 Configuration of ACTIVE Workstation

to serve as new terminals (e.g., Macintosh, NEC-9800 series).

What is notable about the system design is that all its components are organically interconnected with the workstation installed as the system core, permitting real-time access to any subsystem at any time according to the needs of the data analysis.

B. Software Environment

The system currently provides a well-integrated hardware environment, but its software environment requires further sophistication. The system software developed thus far includes: (1) integrated set of data analysis techniques, (2) management software for graphical devices, (3) interface management software, and (4) the handlers and associated software for peripheral devices. The integrated set of data analysis techniques includes the hierarchical classification methods, k-means method, classification methods such as the MST (Minimum Spanning Tree) and multivariate AID methods, principal components analysis, correspondence analysis, simple descriptive statistical methods, etc.

Graphics functions are available for monochrome bitmap displays and color graphics displays. The two color graphics displays are intelligent and have a number of graphics and mathematical primitives included in removable ROM cartridges. The handlers developed to use these ROMs realize plotting and color handling of almost all kinds. Although kinetic display is rather difficult by virtue of the system configuration, it can nevertheless be realized by adequate application of the graphics server processor. Other system components include tablets for linework information input and color XY-plotters for graphics output.

The ACTIVE workstation is controlled from the main workstation via a channel controller which allows any of the subsystems to be connected directly, or indirectly, to any other or to the host system. Analysis data and results, both statistical and graphics, can be freely transferred among the subsystems. Finally, each of the subsystems can be used independently as a standalone system.

IV. CASE STUDIES

In the following sections, some of major studies made using the ACTIVE workstation will be introduced. Although many of these studies have already been reported at some conferences, we consider it worthwhile to review them here as examples of the effective application of computer graphics and automatic classification methods. This is because each of them serves as evidence of our contention that the data analysis that we hope for can never be realized by a mere aggregation of graphical representation techniques.

A. Areal Clustering System and Its Applications

The areal clustering system of the ACTIVE workstation is an integrated set of classification techniques developed for areal partition using multivariate observations that have locate coordinates on the plane (5), (6). These techniques can be broadly classified into two phases according to the features of the data handled. One is NTAP (Numerical Techniques for Areal Partition) and the other is IMAGE (Image Generator of Colored Maps). The preceivable application for NTAP include classification using LANDSAT data and areal partition using land-use data. NTAP is extensively used in

the analysis of data from remotely-sensed measurements.

The objective of IMAGE is to generate image information covering the entire area from discrete and irregularly distributed multivariate observations on the plane (accordingly, on the objective area) for visual presentation of the general features of such partial data and for drawing out the characteristics which are hidden in them. From the partially located data on the screen, IMAGE generates a visual map covering the whole screen and is aimed at stimulating the observer's cognitive powers. As an example of IMAGE applications of the past, we can cite the case where it was used in the analysis of attitude survey data of urban dwellers for the purpose of generating various "colored attitude maps" from the survey results of urban environment assessment. The distribution of attitude data in the survey area, obtained from respondent answers to any specific questions, can be displayed as such colored attitude maps. With a simple glance at these maps, the observer can acquire a firm grasp of respondent opinions to the question, and can also compare the attitude survey data with other kinds of area information, such as amount of greenery. Furthermore, if several questions are asked, the data obtained from respondents can be converted to composite scores by correspondence analysis and principal components analysis and displayed as colored maps. It is also possible to overlay the colored maps of individual questions and to generate composite color maps by image processing.

B. Use of Color Imaging in Automatic Classification

There are many ways of representing multivariate data using computer graphics. New, colored representation methods have been proposed because the raster graphics display has been made much easier to use in recent years (10). Color

image handling is one of the features of our system, and it is possible to develop programs making fuller use of various color handling functions. Development of such programs is made easy by the powerful graphics software and the integrated set of color model transformation programs that we have developed (8). To present the results of various analyses, including automatic classification, immediately on the color display, it is necessary to use a color handling interface program for linking the data analysis techniques and the color monitor. Special attention should be paid to the color models used to simplify the user interface. In other words, the program must be developed using the RGB (Red, Green, and Blue) model for hardware control and the HLS or HSV (Hue, Saturation, and Lightness or Value) models which are closer to the human color perceptional models, for the analyst. If development of this program is to be completed in a short time, it is essential that the color handling system be applied using color models, based on the set of transformation programs between color models. A program for color imaging of automatic classification results is generated using basic software consisting of these programs and the graphics service routines provided as part of the standard system software (9).

Unlike the case where color assignment is made using a color look-up table and a limited number of colors, the program thus generated presents the following features, and problems.

1. Changes in color are controlled on the basis of the three color elements (hue, saturation, and lightness).
2. Accordingly, each color bears a certain meaning and corresponds to a certain tendency of the classification results (colors are not just painted on).

3. However, perfect color handling is still very difficult because of numerous problems remaining in color science. Hence, program development must be planned with a clear understanding of the differences between the color image on the computer graphics display and the human color perception.
4. Peripheral equipment must be used with a clear understanding of the differences between the color system adopted for printing and reproducing color image operations and that adopted for the computer graphics.

Colored Pattern Matrix and Color Patterns

The greater part of the quantitative data handled in data analysis is made up of multivariate data matrices consisting of individuals by variables. The objective here is to represent the features of such a matrix as a color pattern. The color model transformation programs play an important role in the following process required for features representation, although the process itself is quite simple.

(S1) Individuals of the given multivariate data matrix X are classified by a suitable method and reordered.

(S2) Variables are also reordered by a suitable method using, for example, principal component loadings.

(S3) The doubly reordered matrix and each of its elements is converted to color information using the HLS model. The new matrix prepared is taken as "colored pattern matrix" Y, and each of its converted elements is transmitted to the color monitor through the RGB color model to represent the matrix Y as a "color pattern." Next, the lightness is linearly rescaled to reflect the dispersion of distribution of each variable of the multivariate data. The hue and saturation are determined using the component loadings derived from the principal components analysis of the original data matrix X.

Coloring Principal Component Scores

Since the vector corresponding to each individual of colored pattern matrix Y is expressed by means of the three color elements, it is called a "color vector." A plot is prepared that shows the arrangement of scores obtained for any two components derived from the matrix X, and the color vector is placed on the coordinates showing the principal score of each individual. In this way, each of the multivariate observations can be observed as a "color belt" within the reduced dimensions. This method offers the following advantages: (1) ambiguity of principal components scores can be removed, and the meaning of the scores can be shown as color, (2) patterns of clustering can be enhanced in color, (3) discrepancy among clusters and variable features cause by outliers can be grasped with ease in color, and (4) variables characterizing individuals can be compared as changes in color.

The most outstanding advantage of this method is that the features of the original data matrix X can be observed directly as colored images. Dendrograms, scatter diagrams and distribution of principal components scores can naturally observed in time with the obtained color patterns (see the experimental results described below).

C. Classification of Large Data Sets and Its Representation

The main memory size of the workstations and minicomputer of the ACTIVE workstation is not very large. For this reason, considerable difficulty was involved in classifying large-scale data sets (of at least 5,000 to 10,000 cases) or to represent classification results quickly on the display. However, since workstations and minicomputer can be used as

terminals of a large capacity host system, it is certainly possible to classify large-scale data sets using the statistical systems implemented in the host system (SPAD, GENSTAT, SAS, MINTS). The results of such clustering can be transmitted to workstations and minicomputer by file transfer and represented on the display after editing and data modification. Nevertheless, there is no doubt that interactive workstations and minicomputer are more suitable to increasingly detailed and sophisticated data analysis while watching the classification process step by step. Moreover, various ideas are required in order to apply large-scale data clustering to microcomputers with small memory sizes, and it is useful to watch the classification process by graphical representation because the algorithm is often heuristic in nature.

The objective here is not to directly display the distribution of large-scale data. The graphical representation is applied as a means of presenting the skeleton of the data structure using the classification results. In other words, it is applied for the purpose of representing the intrinsic data structure or the fundamental data features on the display. The process actually adopted for classification is as follows.

(S1) A partial data set is generated by sampling a given data set at random in a certain proportion.

(S2) Sampled data is subjected to the initial classification. At this time, the number of clusters is designated to be as large as possible to generate a large number of groups. The distance table is not used because it calls for the use of a large matrix. The k-means method is used in a hierarchical manner to obtain the tree structure.

(S3) The classification thereby completed is displayed by tree representation as the "preliminary classification." Next, a tree showing cluster features is produced using

the fractal recursivity described below.

(S4) All individuals in the data set subjected to the initial classification are classified consecutively using the tree obtained in (S3) as a decision tree (i.e., each individual is assigned consecutively to some cluster in the tree by binary decision rule).

(S5) After all individuals is assigned, "reclassification" is executed using the centroid vector of each cluster at the terminal end of the tree. The reclassification made in this step is the fine adjustment of the relationship between each cluster and the individuals belonging to each cluster. At this time, the tree structure is reconstructed by making use of the reciprocal nearest neighbor (this reconstruction is called "refinement using back-tracking").

(S6) The fractal tree structure obtained by reclassification is displayed and observed again, and is compared with that obtained by the preliminary classification.

(S7) A comparison of any two clusters in the tree is observed by displaying it as a type of Lissajous figure.

Method of Tree Structure Representation

The tree structure obtained in (S5) and (S6) above differs slightly from ordinary dendrograms. To generate this tree representation, attention is directed to the fractal technique of tree models. The fractal plotting is intended to make the plotted tree model look like a natural tree as much as possible. Here, attention is focused on its recursivity and parameter control, with ingenuity exercised so that the tree structure can reflect the data structure to an extent and the computer program can be simplified for faster image generation.

In general, when generating a tree structure, the changes in the tree form are controlled by specifying the following parameters: (1) exterior angle between the limbs, (2) ratio among limb lengths (3) number of limbs, and (4) limb diameter, decline in limb diameter toward the treetop, limb bending, etc.

The tree structure can be made as complex as desired by varying the values of these parameters. Since branching of each limb of a cluster is equivalent to the decomposition of the sum of squares, the following rules are established: (1) the between-cluster sum of squares corresponds to the angle between the limbs, (2) the within-cluster sum of squares (or within-cluster variance) is applied to the length of the limbs, (3) limbs are allocated to the left and right regularly according to the values of the within-cluster sum of squares. This is related to the bend of the limbs.

Graphical Representation of Relationship between Clusters

The tree structure has a drawback to intuitively grasping the similarity between the clusters. Hence, it is necessary to develop a method of observing the relationship between clusters. Since multiple individuals are contained in each cluster and all are multivariate data, the following method was devised.

(S1) Mean vectors of clusters are compared, and the similarity among them is displayed by graphical representation.

(S2) Andrews' functional plotting, a graphical representation method for multivariate observations, is taken and projected to produce a two-dimensional plot.

The p-dimensional mean vector of a cluster is expressed as $\underset{\sim}{x}$, and that of another cluster as $\underset{\sim}{y}$. At this time, the two polynominals obtained by Andrews' method are expressed as follows:

$$f(\underset{\sim}{x}\ ;\ t) = x_1 / \sqrt{2} + x_2 \sin t + x_3 \cos t + x_4 \sin 2t + \ldots$$

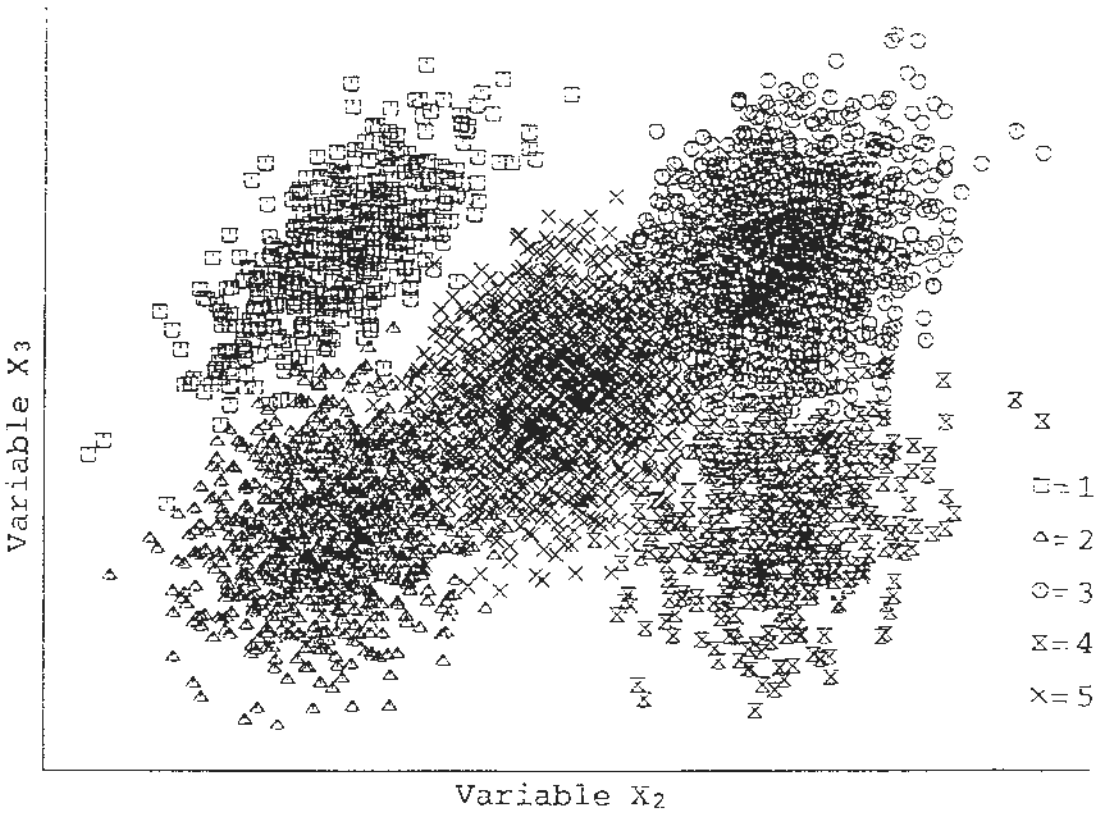
$$f(\underset{\sim}{y}\ ;\ t) = y_1 / \sqrt{2} + y_2 \sin t + y_3 \cos t + y_4 \sin 2t + \ldots$$

where, $-\pi \leq t < \pi$.

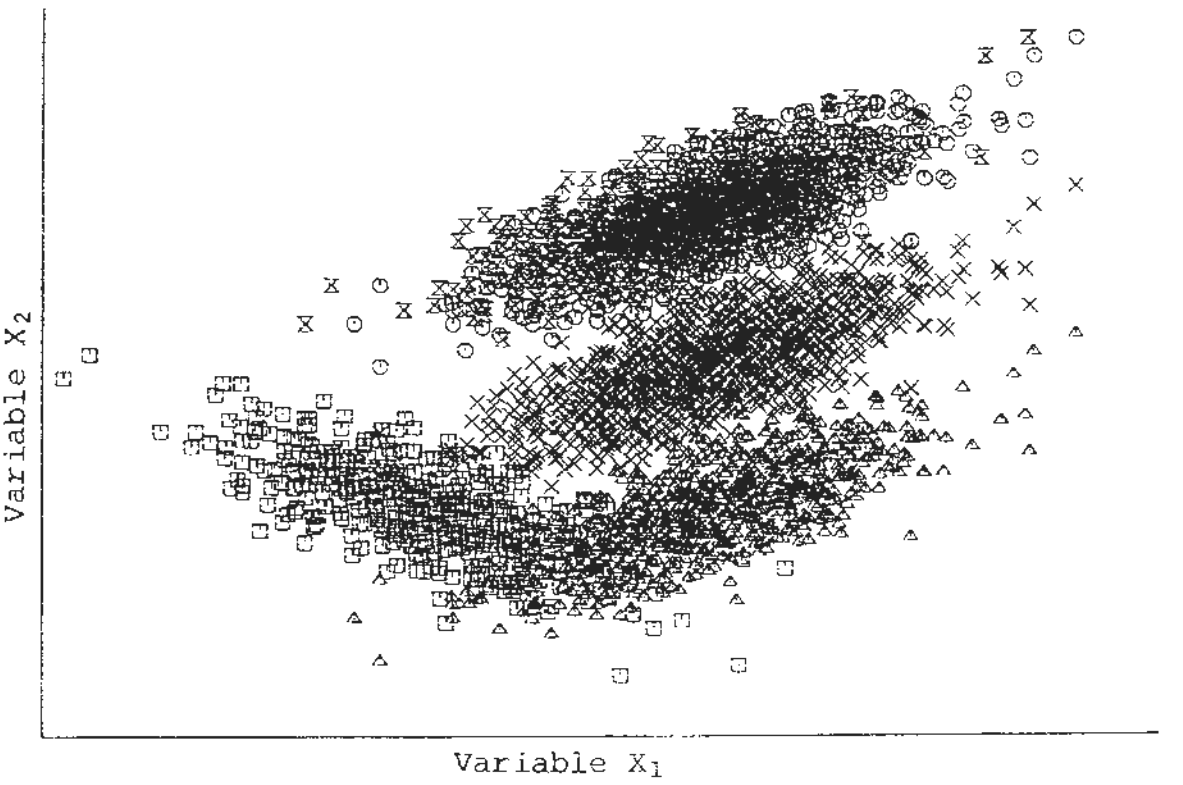
These two functions are projected within the two-dimensional space to obtain the trajectory of the point $(f(\underset{\sim}{x};t), f(\underset{\sim}{y};t))$. The figure plotted by the trajectory clearly shows the similarity between the two clusters, and is a type of Lissajous figure. It is possible to compare mean vectors of clusters by drawing a large number of one-dimensional Andrews' plots. However, observation of the relationship between the two clusters by two-dimensional plotting at each stage of the tree structure in sequence is displayed as the changes in cluster structure, and consequently produces quite different information to that obtained from an overall observation of the tree. Further discussion will be required to explore the mathematical significance of the figures obtained by this method and the relationship between the Lissajous patterns and the data structure. It is generally thought, however, that Lissajous figure can be used as an effective tool of visual comparison of two sets of multivariate observations. It can also be used for graphical representation of similarity or dissimilarity among the individuals.

D. Some Experimental Results

Below are the experimental results obtained using artificial data sets for the purpose of examining the effectiveness of the methods discussed in B. and C. above. The artificial data was prepared by combining five groups of three-dimensional normal random numbers (see reference 2). The number of individuals is 5,000; Figure 2 shows the scatter diagram.

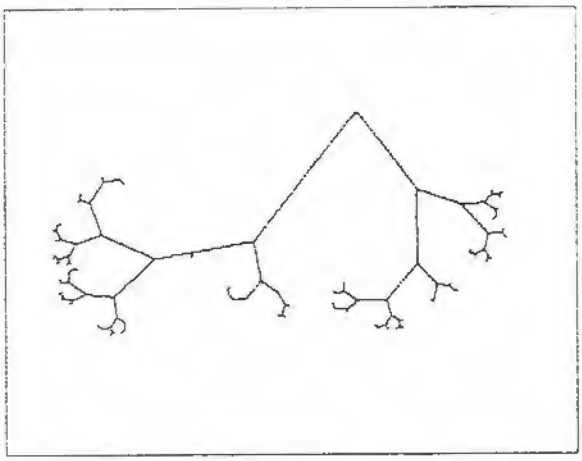


(a)

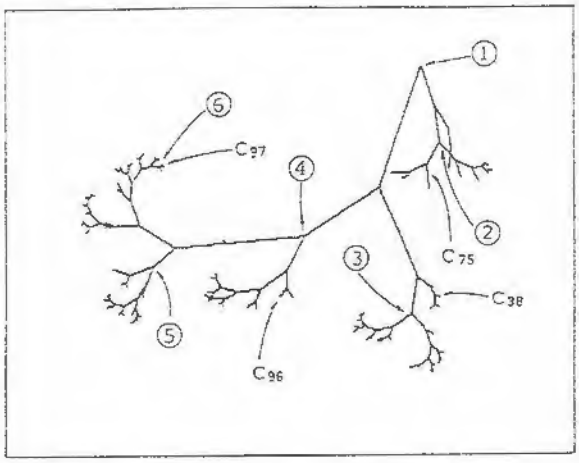


(b)

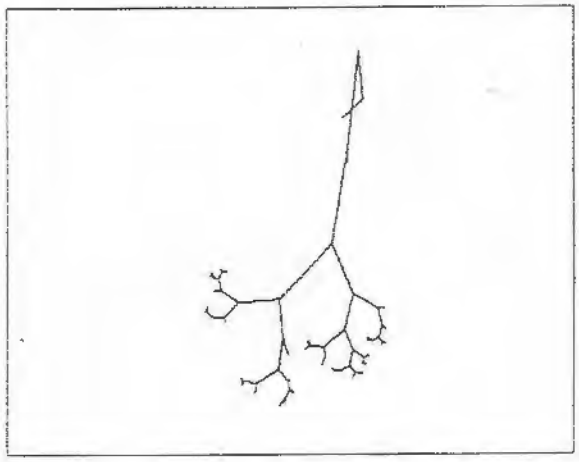
Figure 2 Five thousand observations produced by generating three-dimensional normal random numbers



(a) Result of preliminary classification



(b) Result of reclassification



(c) Another result of reclassification

Figure 3 Tree structures presenting the results of classification

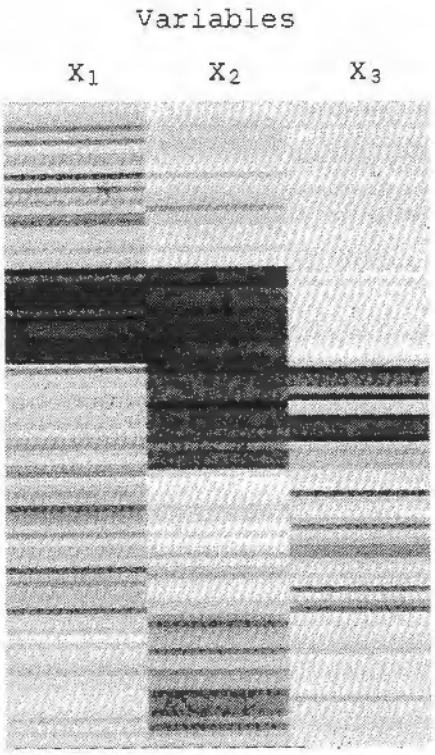


Figure 4 Example of colored pattern matrix

Result 1

Figure 3 (a) and (b) show the tree structures presenting the results of preliminary classification and reclassification of the artificial data. It can be seen that the features of the original data are fairly well exposed by the tree structure, as indicated by the fairly large limb angles. In other words, the outline of the large given data set is enhanced by the tree. For the purpose of comparison, a certain group of data (the group shown in the upper left corner of Figure 2-(a)) was sampled from the given data for similar tree representation shown in Figure 3-(c). Unlike the previous results, the homogeneity of data can be clearly observed from the small angle of limb branching.

Result 2

The feature of clusters obtained in the analysis can be also observed as the color pattern described before (see Figure 4). Although the pattern is printed here as a monochrome copy, it shows the features of each cluster quite well. Compared with the results of Result 1, the dispersion in each cluster is represented more clearly as changes of color in the variables, especially, variable x_2.

Result 3

Comparison among the clusters was performed using Lissajous figures. Figure 5 shows a comparison of two clusters numbered ① to ⑥ , respectively, in Figure 3-(b). Since three-dimensional data was used in this case, the generated curves are rather simple. If data with more dimensions is used, however, complex, yet still useful figures will be generated. The symmetry, inclination, location of

Figure 5 Visualizing the relationships among the clusters

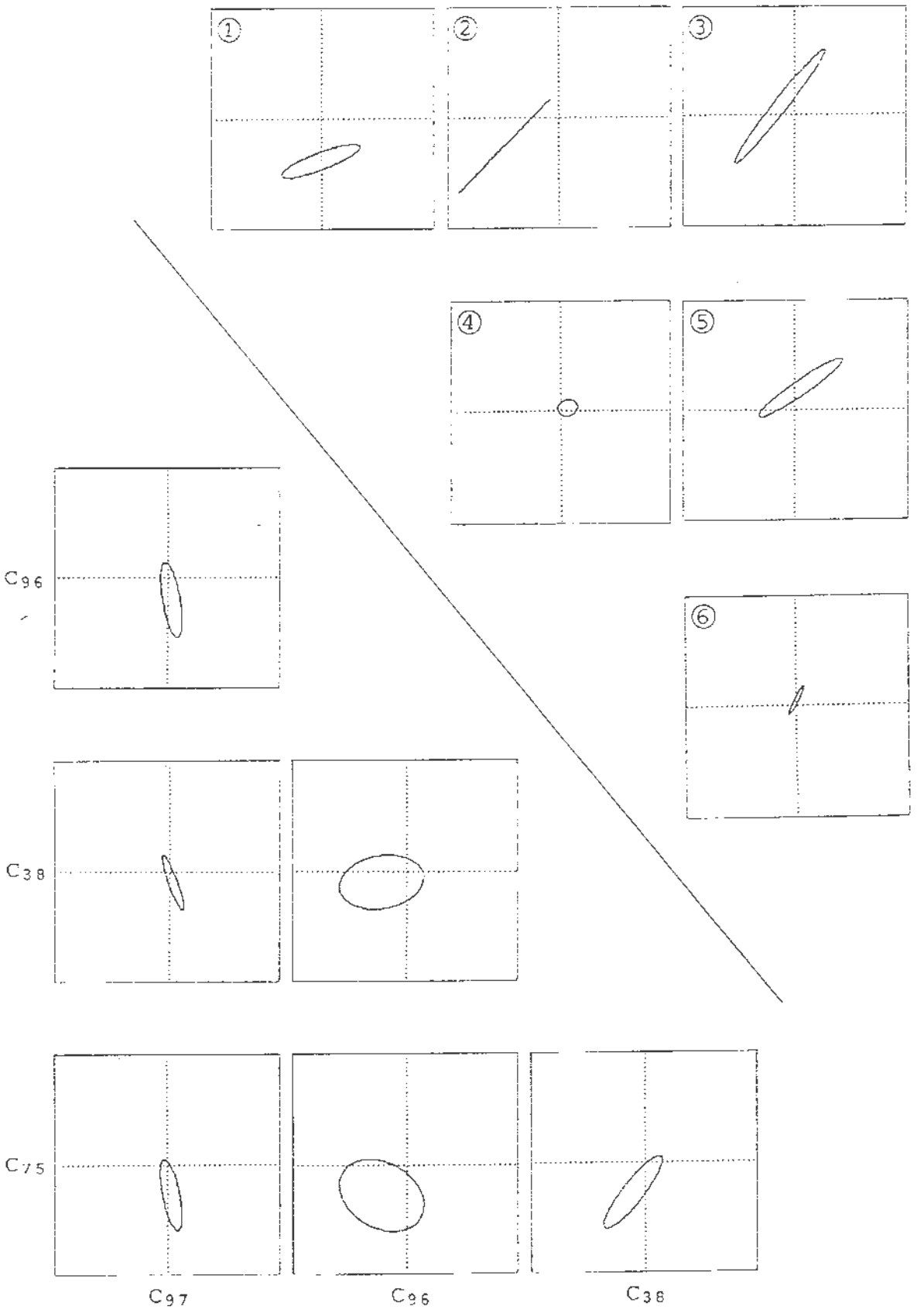


Figure 6 Pair comparison between two clusters

center, convexity of the figures shown in this analysis represent the similarity and difference among the clusters. If we take four clusters from the figure (C_{38}, C_{75}, C_{96} and C_{97}) and draw a pair comparison of their mean vectors as a Lissajous figure, the result is shown in Figure 6. These are quite useful at showing the relationships among the clusters. This is equivalent to plotting the distance table for the clusters. Thus Lissajous figures aid in visualizing cluster relationships which are insufficiently clear in the tree structures.

V. CONCLUSION

There are none who argue against the high utility of graphical representation for quick visual inspection of complex data structures. At the moment, however, reluctance is generally felt in using the information obtained by graphical representation as a direct clue for making decisions in data analysis. The graphical representation methods are constructed rather as a means of grasping intuitive information that can be used in opening up the way to higher sophisticated model analyses. It is precisely for this reason that great importance is attached to the building of a well-integrated computer system environment which is suitable for graphical data analysis. Though small in scale, the ACTIVE workstation introduced in this paper has shown its efficiency as an experimental system. The following improvements are planned for further functional upgrading: (1) development of a programming language with easier-to-use graphics and of object-oriented tools with sufficient extensibility to smoothly link graphics functions using classification results with the algorithms of graphical

representation, and (2) raising the system to the level of an expert system.

ACKNOWLEDGEMENTS

The author is very indebted to Miss M. Koiso who assisted us in the development of several useful computer programs related with section C. in part IV.

REFERENCES

1. Everitt, B. (1978): Graphical Techniques for Multivariate Data, Heinemann Educational Books.
2. Matusita, K. and Ohsumi, N. (1980): A criterion for choosing the number of clusters in cluster analysis, in "Recent Developments in Statistical Inference and Data Analysis," (K. Matusita, ed.) 203-213, North-Holland.
3. Morineau, A. and Lebart, L. (1986): Specific clustering algorithms for large data sets and implementation in SPAD software, in "Classification as a Tool of Research," (W. Gaul and M. Schader eds.), North-Holland.
4. Murtagh, F. (1985): Multidimensional Clustering Algorithms, COMPSTAT Lecture 4, Phisica-Verlag.
5. Ohsumi, N. and Mizuno, K. (1984): Microcomputer graphics clustering techniques for urban environmental evaluations, Proceedings of the XII International Biometric Conference, 148-156.
6. Ohsumi, N. (1984): Practical techniques for areal clustering, in "Data analysis and Informatics III," (E. Diday and others eds.), North-Holland.
7. Ohsumi, N. and Mizuno, K. (1985): Areal clustering system and its application to human ecological problems

(in Japanese), The Proceedings of the Institute of Statistical Mathematics, 33, 2, 176-198.

8. Ohsumi, N. (1986): Transformation programs between color models for computer graphics (in Japanese), The Proceedings of the Institute of Statistical Mathematics, 34, 1, 37-57.

9. Ohsumi, N. (1986): Practical use of color imaging in automatic classification, COMPSTAT-86: Proceedings in Computational Statistics 1986, (F. De Antoni and others eds.), 489-497, Physica-Verlag.

10. Wegman, E. J. and DePriest, D. J. eds. (1986): Statistical Image Processing and Graphics, Marcel Dekker, Inc.