An Application of Statistical Graphs to Cluster Analysis

(1990年４月９日)

Mamoru Fukumori

Key words: Graphical Methods, Cluster Analysis, Dendrogram

Abstract

In order to discuss the effectiveness of graphical methods in cluster analysis, I compared some graphical methods with a dendrogram using some illustrations. This paper shows the use of various graphical methods together with dendrograms. It is indicated clearly that the use of a face graph or a letter graph is very effective in the classification of data.

1. INTRODUCTION

There have been various graphical methods for statistical data analysis proposed since Playfair (1786) proposed a pie graph. Recently, a remarkable improvement in computer graphical faculty has made it possible to draw statistical graphs easily, and very unique statistical graphs for multivariate data have been proposed, such as the face graph (1973), the rader chart, the constellation graph (1973), the linked vector graph (1974), the biplot graphic display (1971) and the Andrews plot (1972).

The main purpose of graphical methods is to make data characteristics, which are normally difficult to understand in the numerical analysis, easy to understand at a glance.

The purpose of this paper is to discuss the effectiveness of statistical graphs in cluster analysis by means of their comparison with dendrograms.

2. GRAPHICAL METHOD FOR CLASSIFICATIONS

Generally, dendrograms are used in cluster analysis. Another approach to the classification of an object is with the use of graphical methods.

Four graphical methods are used in this paper representing multivariate data: the Andrews plot, the face graph, the letter graph and the constellation graph. The following is the outline of each method.

1. Andrews plot

Andrews (1972) suggested transforming a P dimensional response vector x'= $(X_1, X_2, X_3, \cdots, X_p)$ by the Fourier serıes

$$fx(t)=\frac{X_1}{\sqrt{2}}+X_2 \sin t+X_3 \cos t+X_4 \sin 2t+X_5 \cos 2t+\cdots\cdots$$

over the range $-\pi < t < \pi$. Figure 1 shows this graph.

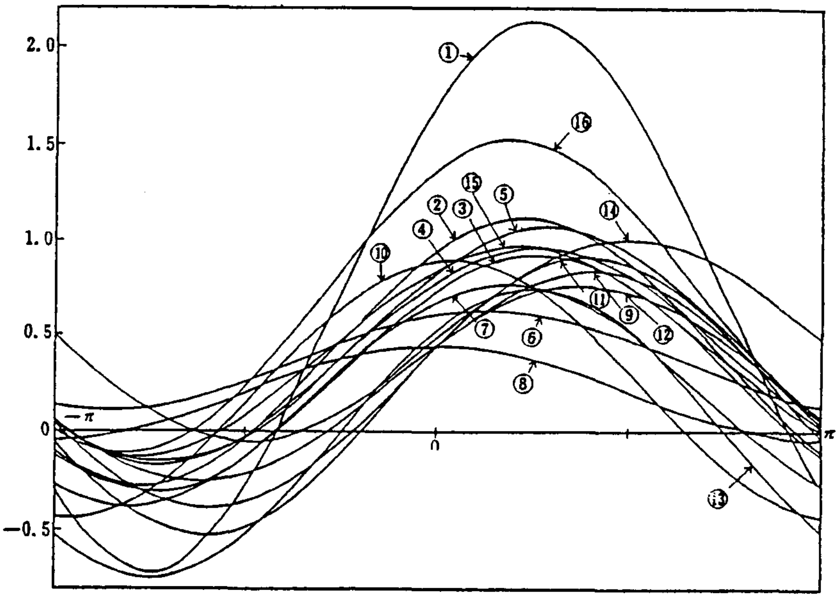

Figure 1 Plot for measurements of air pollution. (Source : Wakimoto (1979)).

Andrews plot has a number of useful statistical properties. Four important properties of the function $f_x(t)$ are:

① The function representation preserves the mean in the sense that if X is the mean of a set of n multivariate observations X_i, then $f_x(t) = (1/n) \Sigma f_{xi}(t)$.

② The function representation preserves distances. It can be proved that the distance between two functions $f_{x1}(t)$ and $f_{x2}(t)$ measured by

$$\left\| f_x(t)-f_y(t) \right\| = \int_{-\pi}^{\pi} [f_x(t)-f_y(t)]^2 dt$$

is proportional to the Euclidean distance between the corresponding points.

The *implication of this property is that close points, in a Euclidean sense, will appear as close functions,* and distant points as distant functions. This is particularly useful when using Andrews plot for uncovering multivariate clusters and outliers.

③ For given t_0, $f_x(t_0)$ is proportional to the length of the projection of the vector $(X_1, X_2, \cdots, X_p)$ on the vector

$$f_1(to) = \left(\frac{1}{\sqrt{2}} ,\ sin\ to,\ cos\ to,\ sin\ 2to,\ cos\ 2to \cdots\cdots \right)$$

Andrews states that the projection on this one-dimensional space may reveal clustering, outlier patterns, or other peculiarities that occur in the subspace and that may be otherwise obscured by other dimensions. The advantage of this plot is that a continuum of such one-dimensional projections is plotted on one graph.

④ The function representation preserve variances. Obviously, if the components of X are uncorrelated with common variance σ^2, then $\mathrm{var}_x(t)$ is $1/2\sigma^2$ if p is odd, and lies between $1/2\sigma^2(p-1)$ and $1/2\sigma^2(p+1)$ if p is even. In this very special case the variability of the plotted function is almost constant across the graph.

2. Face graph

A face graph was proposed by Chernoff in 1973. Chernoff allowed for up to 18 dimensions in a response vector. Each dimension became associated with one of the 18 facial features as Figure 2 and Table 1. The following describes the facial features in greater detail.

[Outline of face] The outline of the face is composed of two ellipses intersecting at points P and P' and is symmetrical with respect to a vertical axis passing the origin point 0. U and L represent the upper and lower vertical limits of the face, that is, the upper part of the face is an ellipse through the point PUP' and lower part of the face is an ellipse through the point PLP'. The distances OU and OL are equal.

[Nose] The nose is a line segment that can be varied in length by hZ_6 with the origin point 0.

[Mouth] The mouth is a circular arc that is centered on the vertical axis and passes through the point P_m. If Z_6 is positive, the mouth is downward, and if Z_6 is negative, the mouth is upward.

[Eyes] The eyes are ellipses oriented about the vertical axis. The variables X_{10}, X_{11}, X_{12}, X_{13}, X_{14}, determine the vertical and horizontal positioning, slant, eccentricity, and size of the eyes.

[Pupil of the eye] The pupil of the eye is represented by a point, whose position is determined by $\pm re(2X_{15}-1)$.

[Eyebrows] The eyebrows are line segments that can be varied in length, slant, and vertical positioning by the variables X_{16}, X_{17}, X_{18}.

Table 1 Program Variables Employed in Constructing a Face

Program Variable	Primary Feature Controlled	Description	Expression
X_1	h^*	Distance from origin O to P	$h^*=1/2(1+X_1)H$
X_2	θ^*	Angle between OP and X-axis	$\theta^*=(2X_2-1)\pi/4$
X_3	h	Half-height of face	$h=1/2(1+X_3)H$
X_4		Eccentricity of upper ellipse	
X_5		Eccentricity of lower ellipse	
X_6		Nose length	
X_7	P_m	Position of center of mouth	$P_m=h\{X_7+(1-X_7)X_6\}$
X_8		Curvature of mouth	
X_9	a_m	Length of mouth	$a_m=X_9(h/1X_81)$ or X_9W_m
X_{10}	Y_e	Height of center of eyes	$Y_e=h\{X_{10}+(1-X_{10})X_6\}$
X_{11}	X_e	Separation of eyes	$X_e=W_e(1+2X_{11})14$
X_{12}	θ	Slant of eyes	$\theta=(2X_{12}-1)\pi15$
X_{13}		Eccentricity of eyes	
X_{14}	L_e	Half-length of eyes	$L_e=X_{14}\min(X_e,W_e-X_e)$
X_{15}		Position of pupils	
X_{16}	Y_b	Height of eyebrow center relative to eyes	$Y_b=2(X_{16}+0.3)L_eX_{13}$
X_{17}	θ^{**}	Angle of eyebrow	$\theta^{**}=\theta+2(1-X_{17})\pi/5$
X_{18}	L_b	Length of eyebrow	$L_b=r_e(2X_{18}+1)/2$

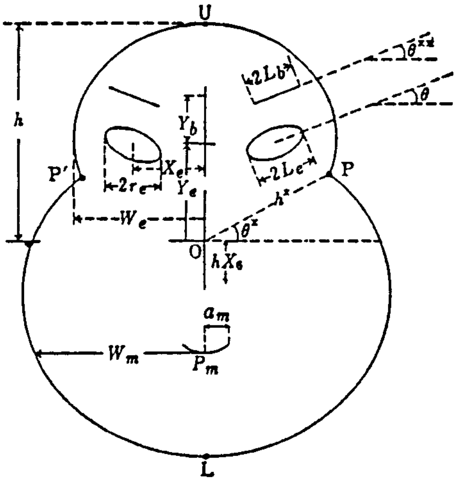

Figure 2 The original Chernoff face

3 . Constellation graph

In a constellation graph, each of the n data with p dimensions is transformed as follows:

$$\begin{cases} \xi_{j\alpha}=f_j(x_{j\alpha})\ ;\ j=1,2,\cdots\cdots,p,\quad \alpha=1,2,\cdots\cdots,n \\ 0\leqq f_j(x_{j\alpha})\leqq\pi \end{cases}$$

$$f_j(x_{j\alpha})=\frac{x_{j\alpha}-x_{jl}}{x_{ju}-x_{jl}}\pi\ ;\ x_{ju}=\max_{1\leqq\alpha\leqq n} x_{j\alpha},\quad x_{jl}=\min_{1\leqq\alpha\leqq n} x_{j\alpha}$$

On each data, the p vectors are linked with each other to form a polygonal line as Figure 3 .

The end of the line determines the position of the star.

As the length of the vector is:

$$\sum_{j=1}^{p} W_j=1\ ;\ W_j\geqq 0,\ j=1,2,\cdots\cdots,p$$

the star is in the semicircle. The position of a star shows the mean and variance of all the variables of

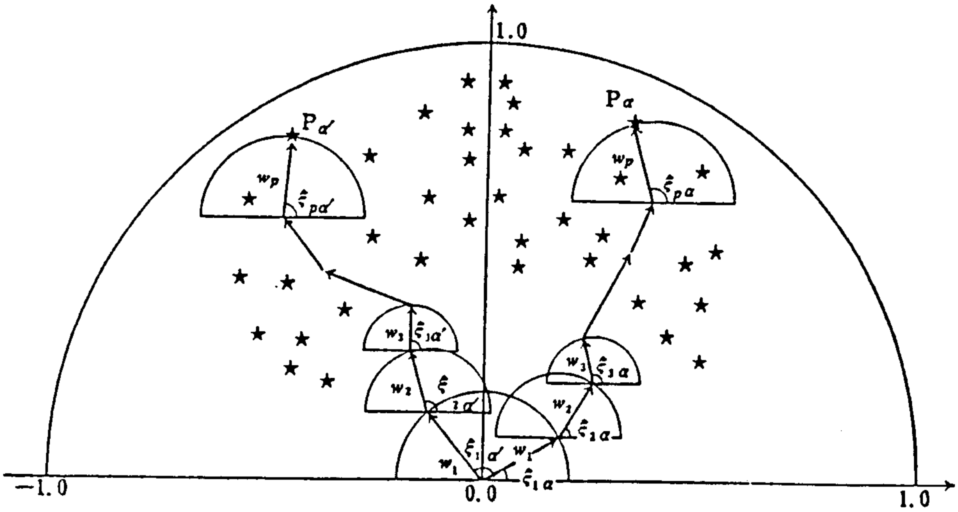

Figure 3 The standard Constellation graph

each subject.

Thus, by means of using the constellation graph, both the feature of a group as a whole and that of individual data are represented simultaneously.

4 . Letter graph

In a letter graph, data value is represented by the size of a letter. That is, each data is put into the length and/or the width of the letter. The drawing procedure of letter graphs is as follows.

First, an appropriate transformation is performed on each data, in accordance with its feature. For example, if each variable has a different unit, the following transformation is done:

$$x_{ij}=\frac{x_{ij}-\overline{x}_j}{S_{xj}}+b\,,\qquad y_{ij}=\frac{y_{ij}-\overline{y}_j}{S_{yj}}+b$$

In this case, b is a positive constant, where X'_{ij}, $Y'_{ij} > 0$. X_i and Y_i are the mean values of variables X_i and Y_i respectively, and S_{xi} and S_{Yi} are the standard deviations of X_i and Y_i respectively.

In the case of examination marks at school, it is suitable to transform them into deviation values as follows:

$$x'_{ij}=10\left(\frac{x_{ij}-\overline{x}_j}{S_{xj}}\right)+50,\quad y'_{ij}=10\left(\frac{y_{ij}-\overline{y}_j}{S_{yj}}\right)+50$$

Next, the letters to be used for this graph are decided, which is the most important step. We have only to assign the value of X_i and Y_i to the letter length and width. In the case of i, the length and width of the first letter are transformed as follows:

$$l\left(\frac{x'_{i1}}{b}\right),\quad l\left(\frac{y'_{i1}}{b}\right)$$

In this case, L is the length of the letter.

Similarly, the second letter is drawn to the right of the first letter in response to the values of X'_{i2} and Y'_{i2}.

The rest of the letters are drawn in the same way.

3 . Application to cluster analysis

Table 2 shows the record of 15 cases with five subjects; Japanese, Mathematics, English, Science and History.

Figure 4 shows a dendrogram drawn by means of Ward method.

The Ward method is based on the loss of information resulting from the grouping of individuals into clusters, as measured by the total sum of squared deviations of every observation from the mean of the cluster to which it belongs.

The assignment rule rests on the increace in the error sum of squares induced from combining every possible pair of clusters. The value, which will be denoted by E.S.S., is used as an objective function. The E.S.S. is computed as follows.

$$E.S.S=\sum_{j=1}^{k}\left(\sum_{i=1}^{nj}X^2_{ij}-\frac{1}{n_j}\left(\sum_{j=1}^{nj}X_{ij}\right)^2\right)$$

Table 2 The score of 15 cases with five subjects

	Japanese	Mathematics	English	Science	History
1	95	87	98	85	80
2	85	90	78	95	54
3	85	70	70	65	85
4	65	70	80	35	25
5	50	40	60	57	60
6	35	45	40	30	40
7	40	95	65	95	50
8	80	60	70	40	30
9	85	80	95	90	80
10	70	85	60	80	50
11	65	70	50	85	60
12	90	85	80	45	40
13	35	45	25	15	20
14	40	70	35	85	30
15	85	90	90	95	90

where X_{ij} denotes the trait value for the ith individual in the jth cluster, k is the total number of clusters at each stage, and n_i is the number of individuals in the jth cluster.

Before I discuss Figure 4, it will be useful to consider more precisely what is represented by a dendrogram.

A dendrogram may be defined as a nested sequence of partitions of the individuals into g groups, where g varies from 1 to n, with the property that the partitions into g and into $(g+1)$ groups are such that $(g-1)$ of the groups are identical while the remaining individuals form one group in the first case and two groups in the second case. Thus a dendrogram is a family of clusters for which any two clusters are either disjointed or one includes the other.

In Figure 4, there appears to be a natural partition into four clusters which have three or four cases.

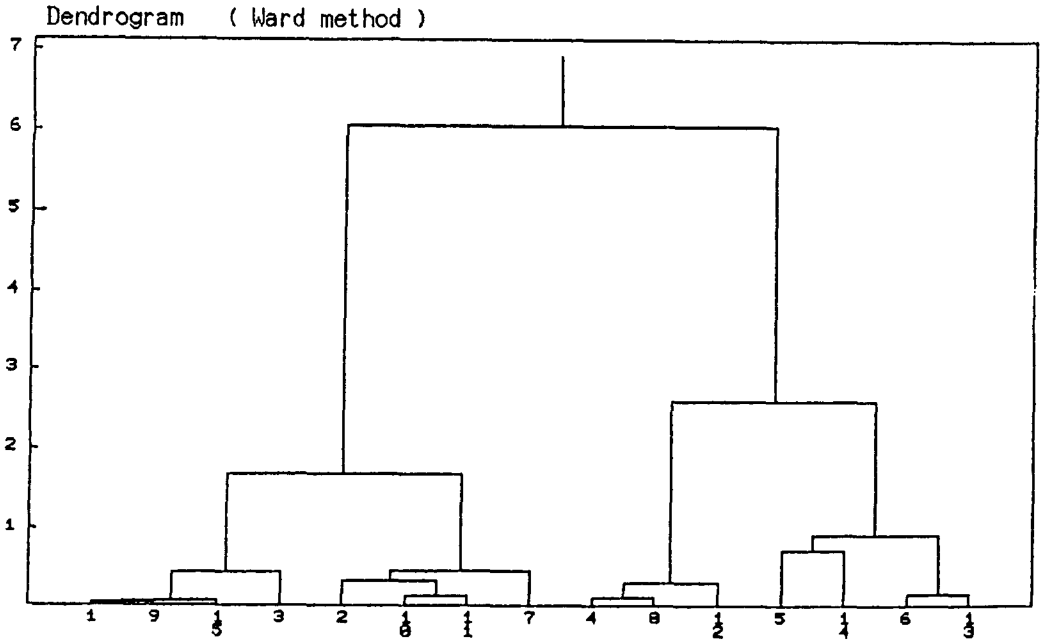

Figure 4 Dendrogram for Ward method on Table 2

Using the same data for Figure 4, the Andrews plot and the Constellation graph can be drawn as Figures 5 and 6 respectively. In comparison with Figure 4, each of the four clusters is not represented in Figure 5.

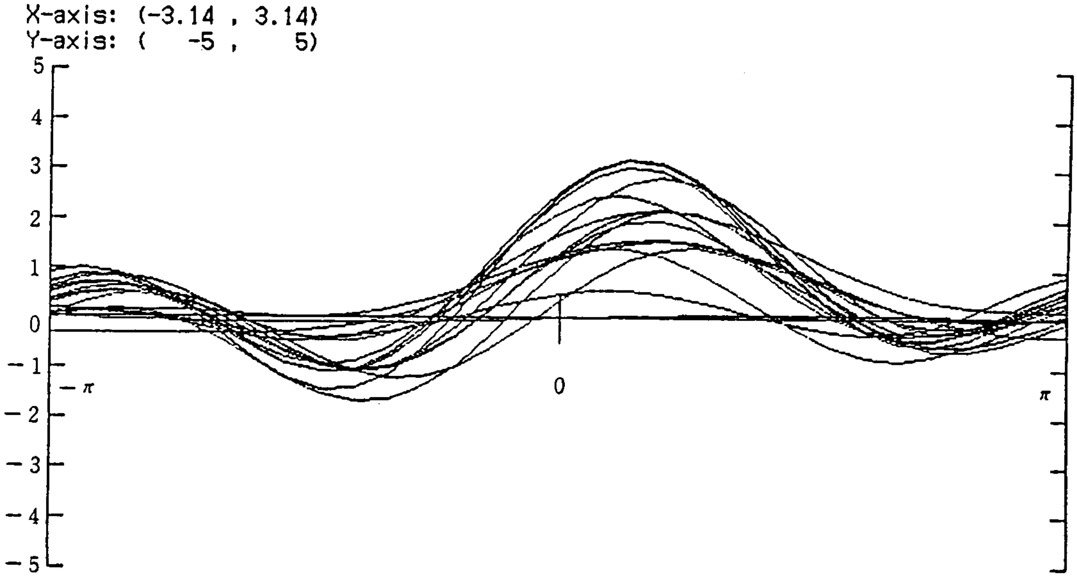

Figure 5 Andrews plot on Table 2

Figure 6 shows a constellation graph. Though the constellation graph has parallel results with the dendrogram in some respects, the four clusters are not very conspicuous.

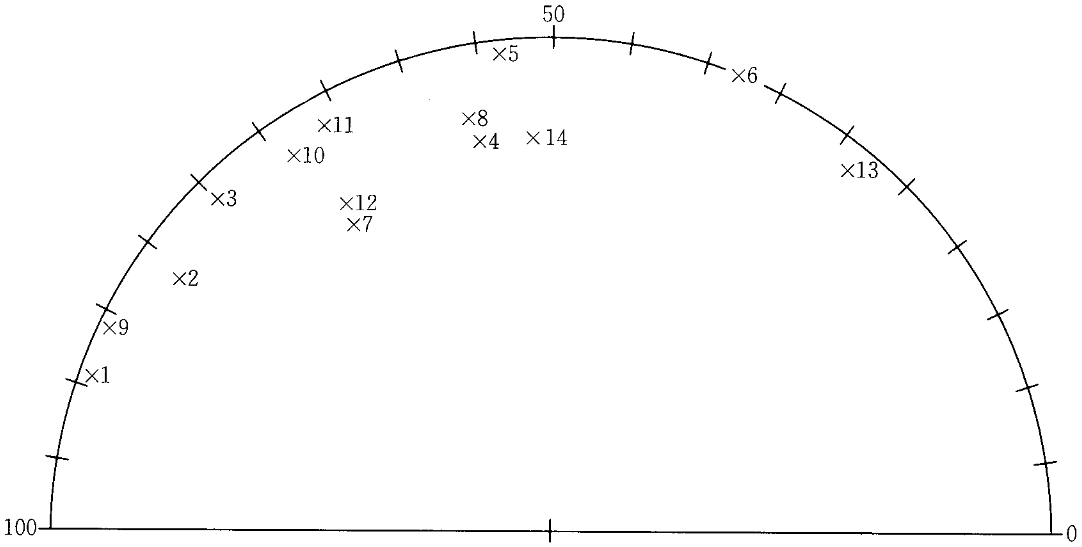

Figure 6 Constellation graph on Table 2

In comparison with the above-mentioned graphs, the next two groups clearly represent the feature of the four clusters.

Figure 7, the face graph, represents each of the four clusters by means of a facial expression. But it is difficult to grasp the feature of each cluster.

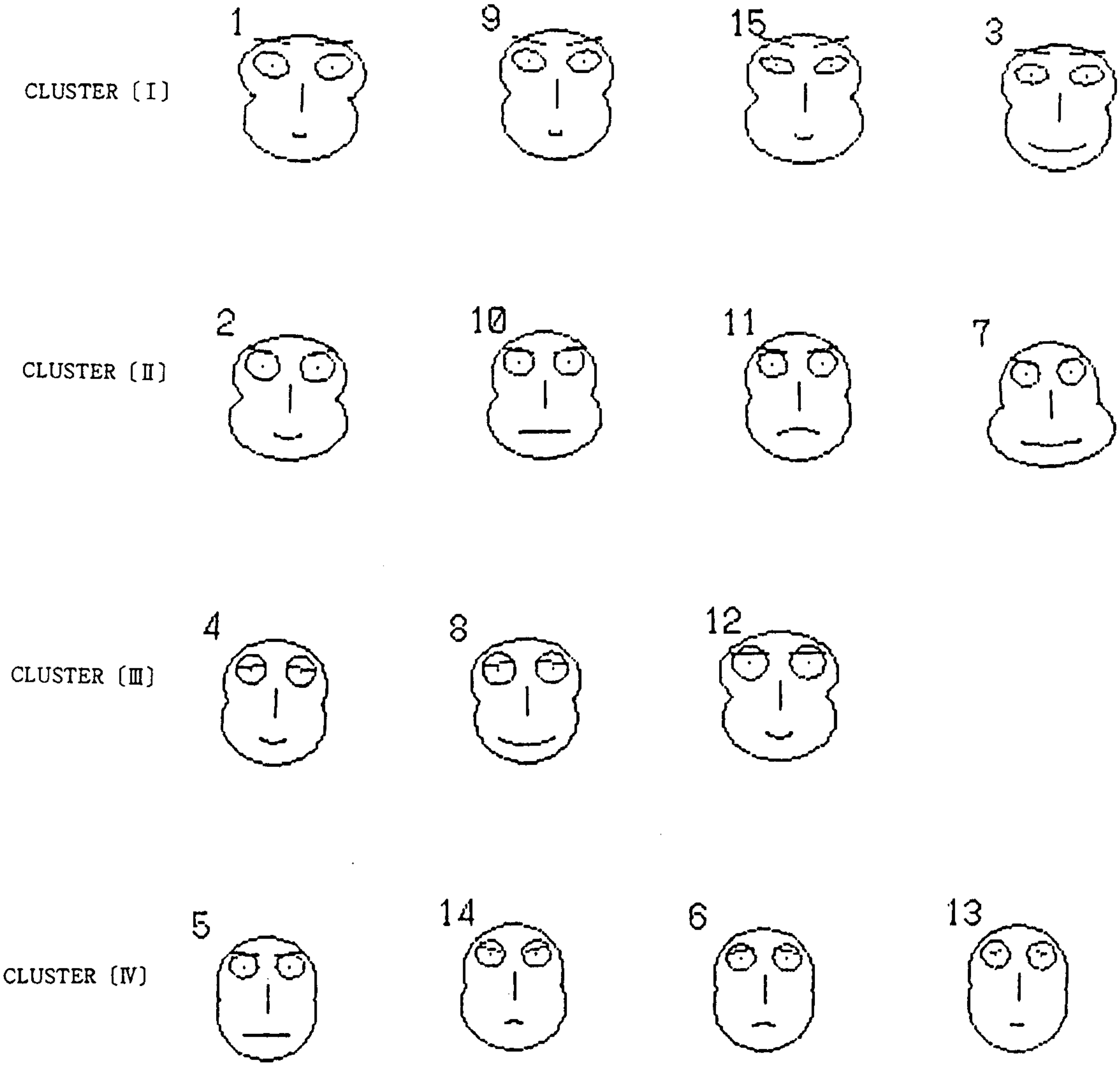

Figure 7　Face graph of four cluster

Figure 8 shows the letter graph. In the letter graph, the property of the four clusters is represented clearly, that is, cluster 1 has the property of relatively higher scores in all subjects, and cluster 2 has the property of Mathematics and Science having higher scores than the three other subjects, and cluster 3 has the property of Japanese , Mathematics and English having higher scores than the two other subjects, and cluster 4 has the property of relatively lower scores in all subjects.

Thus, an application of the graphical method makes it easy to understand the properies of the clusters.

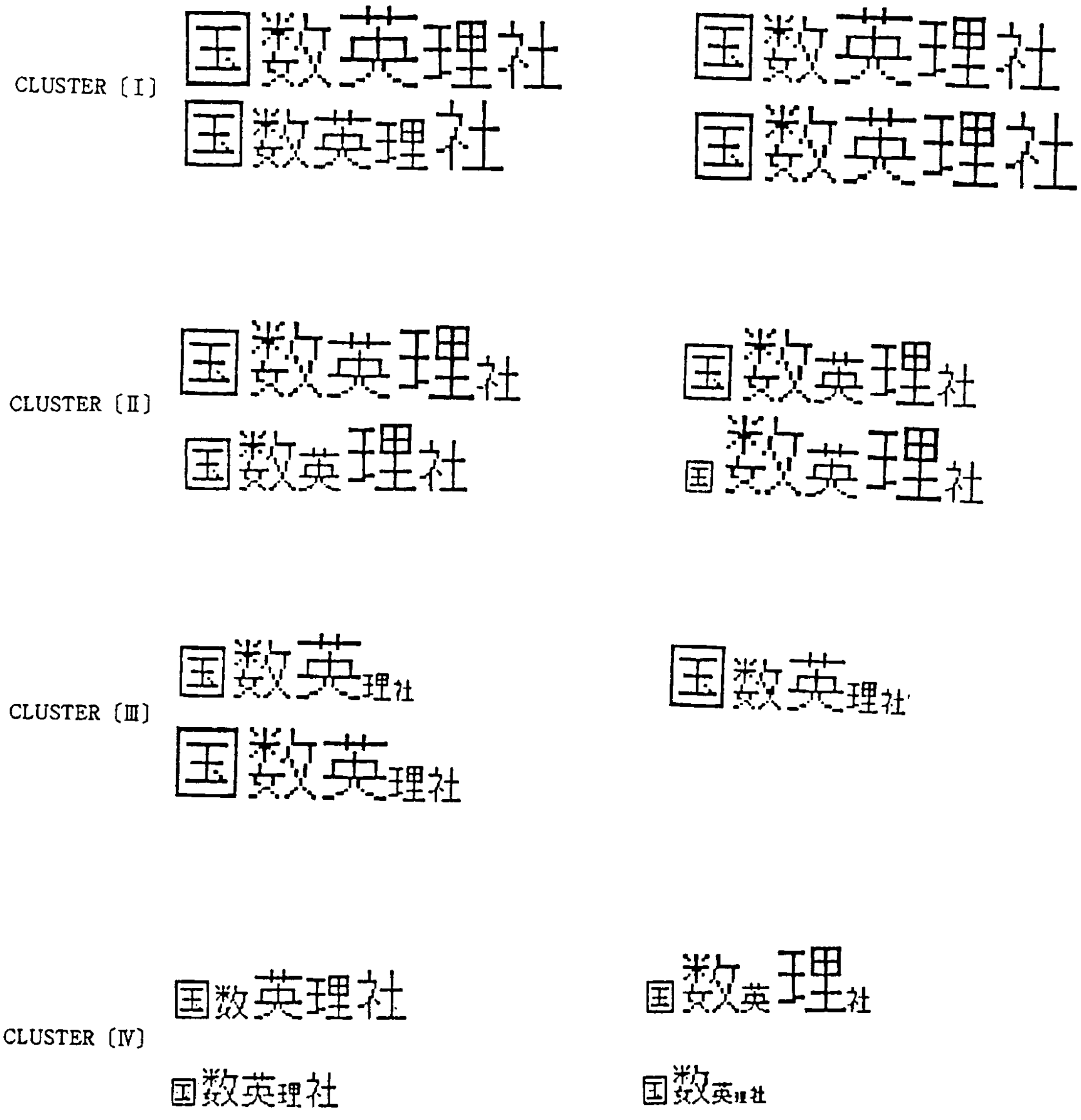

Figure 8 Letter graph of four cluster

4. An idea of using both the dendrogram and the graph method

Generally, a graphical method is used alone. But it is possible to obtain more effective results by means of using many graphic methods together. At the end of this paper, I show an example of using both the dendrogram and the face graph as in Figure 9.

In Figure 9, the process of uniting clusters and the features of the united clusters are clearly shown.

As mentioned above, graphical methods can be very effective in cluster analysis when used properly.

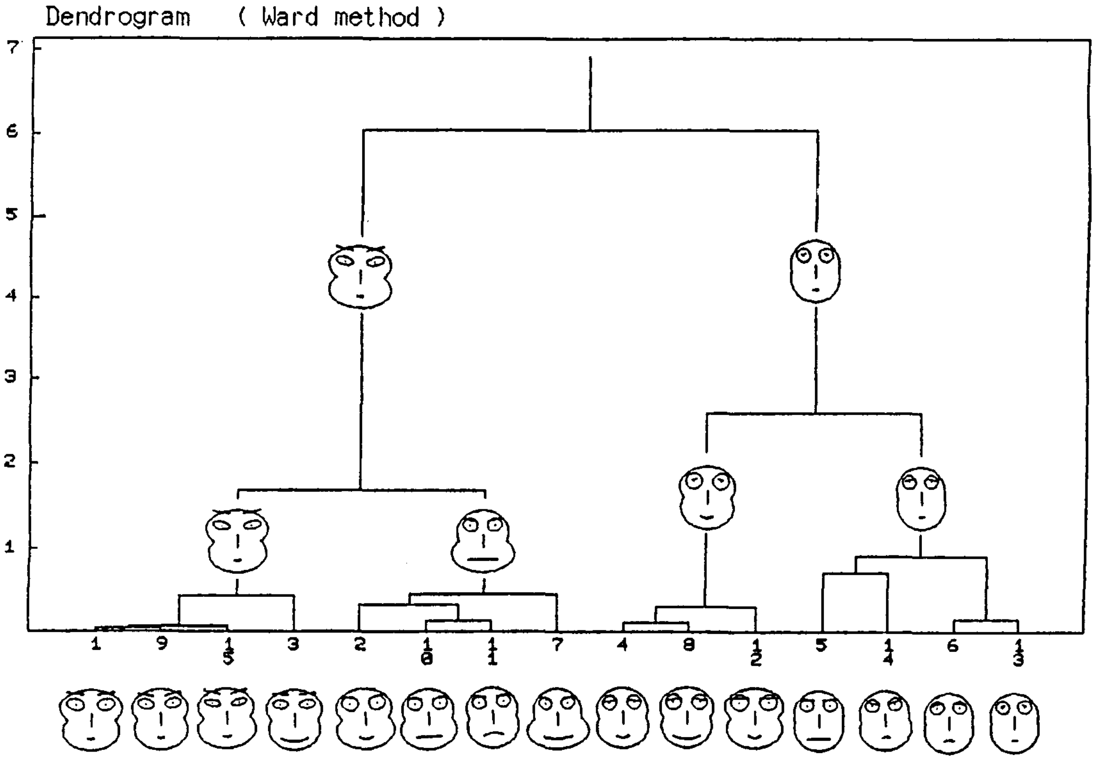

Figure 9 Representation using both the dendrogram and the face graph

References

1) Andrews, D.F. (1972) . Plots of high-dimensional data. Biometrics. 28, 125-136.

2) Chernoff, H. (1973) . The use of faces to represent points in k-dimensional space graphically. Journal of American Statistic Association. 68, 361-368.

3) Corsten, L.C. & Gabriel, R.K. (1976) . Graphical exploration in comparing variance matrices. Biometrics. 32, 851-863.

4) Dillon, W. & Goldstein, M. (1984) . Multivariate Analysis. John Willey & Sons.

5) Gabriel, R.K. (1971) . The biplot graphic display of matrices with application to principal component analysis. Biometrika. 58, 453-467.

6) Hirai, Y., Fukumiri, M. & Wakimoto K. (1988) . Kanji graph representation for multivariate data and its application to cluster analysis.

Bulletin of The Computational Statistics of Japan. 1 , 1 ,11-21.

7) Peter, C.W. (1978) . Graphical Representation of Multivariate Data. Academic press.

8) Toit, S.H.C., Steyn, A.G.W. & Stumpf, R.H. (1986) . Graphical Exploratory Data Analysis. Springer-Verlag.

9) Wakimoto, K. & Taguri, M. (1978) . Constellation graphical method for representing multidimensional data. Annual Institute of Statistical Mathematics.