# WORKING PAPER

## CLUSTER ANALYSIS AS A TOOL OF INTERPRETATION OF COMPLEX SYSTEMS

S. Miyamoto

January 1987

WP-87-41

IIASA
International Institute
for Applied Systems Analysis

NOT FOR QUOTATION
WITHOUT PERMISSION
OF THE AUTHOR

# CLUSTER ANALYSIS AS A TOOL OF INTERPRETATION OF COMPLEX SYSTEMS

S. Miyamoto

January 1987
WP-87-41

*Working Papers* are interim reports on work of the International Institute for Applied Systems Analysis and have received only limited review. Views or opinions expressed herein do not necessarily represent those of the Institute or of its National Member Organizations.

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
A-2361 Laxenburg, Austria

# FOREWORD

This paper deals with several problems in cluster analysis. It appears that the suggested solutions have not been considered in current literature. First, the author proposes the use of a permutated matrix as a tool for interpretation of clusters generated by hierarchical agglomerative clustering algorithms. Second, a new method of defining similarity between a pair of clusters is shown. This method leads to a new class of hierarchical agglomerative clustering. Third, two criteria are defined to optimize dendrograms that are outputs of hierarchical clustering.

This paper has been presented at the Task Force Seminar Session on New Advances in Decision Support Systems, Laxenburg, Austria, November 3-5, 1986.

Alexander B. Kurzhanski
Chairman,
System and Decision Sciences Program

# Cluster Analysis as a Tool of Interpretation of Complex Systems

S. Miyamoto

## 1. Introduction

Recently techniques of the cluster analysis has become a standard tool for analyzing and recognizing objects to be studied in various fields of sciences. One remarkable characteristic of the cluster analysis is that it directly generates several categories of objects without any predefined standards for the classification. Application of the technique of the cluster analysis is easy, since many algorithms of the cluster analysis do not require prerequisites in advanced mathematics. On the other hand, it has been suggested that the cluster analysis has its inherent weak point: it has many algorithms with various options that one can not judge which is the best for a particular application. In many cases, however, this weak point is due to a fundamental property of human psychology in the sense that in natural psychological classification boundaries of categories are not clear, and also categories have hierarchical structure of supercategories and subcategories. Therefore in general we can not solve theoretically the problem to overcome the above weak point and we do not touch this problem in this paper.

In spite of this drawback, experiences in many fields exhibit that the cluster analysis is a useful technique to find structures in a complex system. If we describe data analysis in a very general term as a process starting from a chaos of huge data and disorder of various information to a final goal of clear understanding of system structure with structured configuration of information and with summarized representation of data, the cluster analysis is particularly useful in an early stage of data analysis. That is, the generated structures by the cluster analysis will help system analyst to proceed his analysis by summarizing data and information; in the later stages he should check or varidate the generated categories by some other means including his own knowledge of the system.

This paper does not aim at introducing a new framework of the clustrer analysis, nor is it a survey of the various techniques. We will describe here some problems in the current methods of the cluster analysis together with solutions to them. The aim here is to improve the current techniques for better application to real problems and to show some ideas that will be important in future studies of the cluster analysis.

In the present paper we are concerned with the hierarchical methods of the cluster analysis, since in many real problems it is difficult to determine beforehand the number of categories to be generated. Note that nonhierarchical algorithms require specification of the number of clusters. If we have sufficient prior information on the number and properties of the categories, various nonhierarchical procedures might be effective, but we do not assume that we already know the number of categories beforehand.

Section 2 deals with a technique of simultaneous clustering of objects and attributes. Section 3 is devoted to some new algorithms of hierarchical clustering that current literature does not deal with. Section 4 shows a method of "optimizing" the output from the hierarchical cluster analysis.

## 2. Twoway clustering

### 2.1 Need for twoway clustering

Let $X=\{x_1,x_2,...,x_m\}$ be a set of objects or entities to be classified. On the other hand let $Y=\{y_1,y_2,...,y_n\}$ be a set whose members are called attributes or variables. Relations between an entity $x_i$ and $y_j$ is described by a real number $c_{ij}$. Therefore we assume that a matrix $C=(c_{ij})$ is given.

Since we consider the hierarchical cluster analysis, (and in particular, agglomerative hierarchical cluster analysis. In the below the word of cluster analysis means agglomerative hierarchical cluster analysis. Exceptions will be written explicitly.) first we should describe the major outline of the hierarchical clustering. Namely, hierarchical clustrering consists of the following two steps:

1. Definition of a similarity measure $s(x_i,x_j)$ between an arbitrary pair of entities $x_i$ and $x_j$.
2. Generation of clusters based on the similarity measure $s(x_i,x_j)$.

In the first stage the definition of a similarity $s(x_i,x_j)$ is based on two vectors $(c_{ik})$, $(c_{jk})$, $k=1,2,..,n$. In other words the space Y is used to define the similarity measure through the matrix C. Various similarity measures have been proposed. We do not describe them in detail. (See Anderberg, 1973.) Therefore we simply assume that $s(x_i,x_j)$ is given by any method for definition of the similarity.

In the second stage there also exists a number of algorithms for hierarchical clustering. In this section we need not describe them. (See Anderberg, 1973; Everitt, 1980.) In general input to a hierarchical algorithm is a matrix $(s_{ij}=s(x_i,x_j))$ of similarity defined in the first stage and its output is a tree-like figure called dendrogram. The output of the dendrogram has great amount of information, since it shows not only the generated clusters but also the procedure of forming clusters one by one. The significance of the dendrogram will be emphasized throughout the whole sections.

Let us recall that X means entities to be clustered and Y means variables that are used to define a similarity. This distinction is, however, for convenience's sake. In practical situation sometimes we wish to cluster Y using X. Furthermore it frequently occurs that we wish to cluster both X and Y. Hartigan (1975) called this as a simultaneous clustering. He proposed a particular method of the simultaneous clustering and it was implemented on BMDP program package. Basically there is no great difference in the method of Hartigan and our method which will be described here. Hartigan's method is, however, too restrictive for applying it to many of real problems. Therefore it is necessary to describe here a method that is similar to Hartigan's method but different from it from a practical viewpoint. In the below we call our method as a twoway cluster analysis.

## 2.2 Pattern in the plane

Let us begin with a simple example. Consider the matrix $C=(c_{ij})$, $i=1,..,5$; $j=1,...,4$ with $X=\{x_1,...,x_5\}$ and $Y=\{y_1,...,y_4\}$ in Fig.1a. We wish to cluster X based on C. Here we assume that $c_{ij}$ means whether $y_j$ is applicable to $x_i$ ($c_{ij}=1$) or not ($c_{ij}=0$). By any definition of the similarity and algorithm, we can obtain three clusters in X shown in Fig. 1b. In practice it is important to see why these clusters have been generated in relation to the set Y through C. One of the best way to see the relationship is to cluster Y as in Fig. 1c.

| | $y_1$ | $y_2$ | $y_3$ | $y_4$ |
|---|---|---|---|---|
| $x_1$ | 1 | 0 | 0 | 1 |
| $x_2$ | 0 | 1 | 0 | 0 |
| $x_3$ | 0 | 0 | 1 | 0 |
| $x_4$ | 1 | 0 | 0 | 1 |
| $x_5$ | 0 | 0 | 1 | 0 |

Fig.1a

| | $y_1$ | $y_2$ | $y_3$ | $y_4$ |
|---|---|---|---|---|
| $x_1$ | 1 | 0 | 0 | 1 |
| $x_4$ | 1 | 0 | 0 | 1 |
| $x_3$ | 0 | 0 | 1 | 0 |
| $x_5$ | 0 | 0 | 1 | 0 |
| $x_2$ | 0 | 1 | 0 | 0 |

Fig.1b

| | $y_1$ | $y_4$ | $y_3$ | $y_2$ |
|---|---|---|---|---|
| $x_1$ | 1 | 1 | 0 | 0 |
| $x_2$ | 1 | 1 | 0 | 0 |
| $x_3$ | 0 | 0 | 1 | 0 |
| $x_5$ | 0 | 0 | 1 | 0 |
| $x_2$ | 0 | 0 | 0 | 1 |

Fig.1c

This simple example shows the importance of the twoway analysis even if we wish to classify only the elements in X. When we use cluster analysis to understand structure of a complex system, what is important is to understand the meaning of the clusters, in other words, to interprete the clusters. The interpretations are given in terms of the relation of a cluster to variables in Y through the matrix C. In general it is more desirable to observe its relation to "a cluster of variables" rather than individual variables.

To realize the above idea in an actual computer program of the cluster analysis, let us define a permutated matrix. Let $\sigma$ and $\tau$ be two permutations of $\{1,2,\ldots,m\}$ and $\{1,2,\ldots,n\}$, respectively. According to orders $(x_{\sigma(1)},\ldots,x_{\sigma(m)})$ and $(y_{\tau(1)},\ldots,y_{\tau(n)})$ a permutated matrix $(\ c_{\sigma(i)\tau(j)}\ )$ is determined. To obtain the permutated matrix that reflects the categorical structures generated by a clustering algorithm, we simply use the orders in the entries of the dendrograms of X and Y. Namely, the outline of the twoway clustering is as follows.

1. Calculate similarity $s(x_i,x_j)$ defined on X x X and $s'(y_p,y_q)$ on Y x Y. Use the same kind of measures for s and s'. For example, in our former studies (Miyamoto and Nakayama, 1986) we proposed the following similarity measure based on the framework of fuzzy sets.

$$s(x_i,x_j) = \frac{\sum_k \min(c_{ik},c_{jk})}{\sum_k \max(c_{ik},c_{jk})} \qquad (1)$$

$$s'(y_p,y_q) = \frac{\sum_k \min(c_{kp},c_{kq})}{\sum_k \max(c_{kp},c_{kq})}$$

2. Perform the clustering by a hierarchical algorithm and output

two dendrograms in which entries are ordered as $(x_{\sigma(1)},\ldots,x_{\sigma(m)})$ and $(y_{\tau(1)},\ldots,y_{\tau(n)})$. Then output the permutated matrix $( c_{\sigma(i)\tau(j)} )$ as a two dimensional pattern of relation between X and Y.

Remark In case that $c_{ij}$ is binary or it means the frequency of occurrence of $x_i$ at $y_j$ (Miyamoto and Nakayama, 1986), zero entries in the permutated matrix should be replaced by blanks so that we can observe the two dimensional pattern more clearly. []

## 2.3 Scaling of clusters

Let us assume that we already have clusters (subgroups) $X_1,\ldots,X_K$ of X and $Y_1,\ldots,Y_L$ of Y by some hierarchical or nonhierarchical clustering. The problem here is to find some scales on $\{X_p\}$ and $\{Y_q\}$, $p=1,..,K$; $q=1,..,L$, so that the resulting display of $X_p$'s on X-axis and $Y_q$'s on Y-axis shows relation between clusters of X and those of Y. This problem has been studied for a long time as optimal scaling problem. This problem in our context can be solved if we can define an aggregated matrix $(\bar{c}_{pq})$ between $X_p$ and $Y_q$ from $c_{ij}$ in a reasonable way. In many cases it is natural to define

$$\bar{c}_{pq} = \sum_{\substack{x_i \in X_p \\ y_j \in Y_q}} c_{ij}$$

Then, the scaling problem can be formulated as follows (See Kendall and Stuart, 1973) Let $\alpha_1,\ldots,\alpha_K$ be coordinates on X-axis to be determined for $X_1,\ldots,X_K$, and $\beta_1,\ldots,\beta_L$ be coordinates on Y-axis to be determined for $Y_1,\ldots,Y_L$. Then we consider maximization of the following criterion:

$$\frac{\sum_p \sum_q \bar{c}_{pq} \alpha_p \beta_q}{\sqrt{\sum_p \sum_q \bar{c}_{pq} \alpha_p^2} \sqrt{\sum_p \sum_q \bar{c}_{pq} \beta_q^2}}$$

The maximization problem is equivalent to an eigenvalue problem (Kendall and Stuart, 1973):

$$\sum_q \bar{c}_{pq} \beta_q - \lambda \sum_q \bar{c}_{pq} \alpha_p = 0 , \qquad p=1,2,\ldots,K$$

$$\sum_p \bar{c}_{pq} \alpha_p - \mu \sum_p \bar{c}_{pq} \beta_q = 0 , \qquad q=1,2,\ldots,L$$

In case that we use hierarchical clustering, clusters $\{X_p\}$ and $\{Y_q\}$ can be obtained by cutting the two dendrograms at a certain levels of similarities.

## 3. A new class of hierarchical clustering algorithms

### 3.1 Similarity between two clusters

A major part of clustering algorithms is based on calculation on a Euclid space. In case of the hierarchical clustering, we have the centroid method and the Ward method. In nonhierarchical clustering the Euclid space has been assumed in general. Algorithms that do not assume a Euclid space is exceptional. This tendency is due to the fact that various good properties of the Euclid space are available to have an advanced algorithms.

For example, in the Euclid space minimization of dispersion within clusters is equivalent to maximization of dispersion between clusters. (See Everitt, 1980.) Nevertheless, in many cases we can not assume the Euclid model in real problems. Even when the Euclid space is assumed, many authors emphasize the necessity of validation of clusters (See e.g., Bezdek, 1981.) based on some criterion that is not based on the Euclid model. In these cases the authors implicitly assume two different criteria: one to generate clusters and the other to validate the generated clusters. Here arises a question: what if we use the latter criterion from the first to generate clusters? In many cases it is possible to use the criterion of validating clusters for the purpose of defining similarity between a pair of individuals.

When we do not assume the Euclid space, available algorithms are far poorer than those based on the Euclid model. For example, in hierarchical algorithms, the single linkage, the complete linkage, and the average linkage methods are available. In these algorithms, however, calculation of similarity between a pair of groups is not based on the initial mathematical model, but is based on a rather simple arithmetic calculation. For example, in the single likage clustering similarity s(A,B) between a group $A=\{a_1,a_2,\ldots\}$ and $B=\{b_1,b_2,\ldots\}$ is defined as

$$s(A,B) = \min_{\substack{a_i \varepsilon A \\ b_j \varepsilon B}} s(a_i,b_j)$$

This observation leads us to development of a new class of clustering algorithms that has various implications for future studies.

Let us consider again the above definition of s(A,B), where s(A,B) is defined only in terms of $s(a_i,b_j)$ as similarity between a pair of individuals. The latter is based on a specific mathematical model, whereas the former is not. Nevertheless, a similarity between two groups can be defined in a natural way that is different from the arithmetic definition of s(A,B) such as the one defined above.

Let us consider a specific example. Consider the model (X,Y,C) in the previous section. We assume here that $c_{ij} \geq 0$ is a frequency of occurrence of $x_i$ at $y_j$ and also the measure is defined by (1). If we consider similarity $s(X_p,X_q)$ between two groups $X_p$ and $X_q$, it is natural to define

$$s(X_p,X_q) = \frac{\sum_k \min(\bar{c}_{pk},\bar{c}_{qk})}{\sum_k \max(\bar{c}_{pk},\bar{c}_{qk})} \qquad (2)$$

where

$$\bar{c}_{\alpha k} = \sum_{x_j \varepsilon X_\alpha} c_{jk}, \qquad \alpha = p,q .$$

Of course other measures such as the cosine correlation are applicable to define $s(x_i,x_j)$ and $s(X_p,X_q)$ in the same way as above. It should be noted that in the latter case the similarity between two groups are defined based on the same model as the one on which the similarity between a pair of individuals has been defined.

The latter definition of similarity between two groups can be used in two ways:

1. Development of a new class of algorithms.

2. Use of similarity between two groups for other purpose than the generation of clusters.

### 3.2 Clustering algorithms

As is mentioned in the previous section, we assume that the similarity s(A,B) between a pair of groups is defined on the same model as the one on which the similarity between two individuals are defined. This means that some aggregation procedure like (2) is defined in a natural way.

When we do not assume the Euclid space, maximization of similarity within groups is not equivalent to minimization of similarity among groups. Therefore in the clustering algorithms we have two different approaches:

1. Generate clusters so that sum of similarities within clusters is maximized.

2. Generate clusters so that sum of similarities between every pair of clusters is minimized.

These two rules are applicable to hierarchical agglomerative clustering, hierarchical divisive clustering, and to nonhierarchical clustering with a fixed number of clusters. In case of hierarchical agglomerative clustering, the former rule is easier to apply. In the following algorithms we assume the set $X=\{x_1,...,x_m\}$ with elements $x_i$'s. Clusters are denoted as $X_p$'s.

(Outline of a hierarchical agglomerative algorithms)

1. Let N:=n (number of clusters) and let $X_i=\{x_i\}$ for all i=1,..,N

2. For all $1\leq i,j\leq N$, $i\neq j$, calculate similarities $s(X_i,X_j)$ based on the given model.

3. Find a pair $X_p$, $X_q$ such that $s(X_p,X_q) = \max_{i,j} s(X_i,X_j)$ and merge them into $X_r = X_p \cup X_q$ .

4. N:=N-1. If N=1 output the result as the dendrogram and stop.

5. For all $1\leq i\leq N$, $i\neq r$, recalculate similarities $s(X_r,X_i)$ based on the same model. Go to Step 3. []

### 3.3 Use of similarity between a pair of groups

The similarity measure s(A,B) between a pair of groups can be used in a number of other ways than the generation of clusters above mentioned. Here we discuss multidimensional scaling of clusters and a method of classification based on the mathematical model of clustering.

Multidimensional scaling has been studied in the field of mathematical psychology (Kruskal, 1964). It projects the data points onto plane or three-dimensional space so that the resulting configuration shows overall structure of the data set. The projection is based on an optimization in the sense that the distance between every pair of elements on the plane (or three-dimensional space) reflects in an optimal way the original similarity defined on that pair. The multidimensional scaling has been used in much the similar way as the factor analysis. Unfortunately in the presence of many points to be projected, for example, one hundred points, the multidimensional scaling often fails: it is more suitable for small number of elements (e.g., 10 - 20 ). In such a case an effective way to apply the method of the scaling is to summarize elements into a smaller number of clusters. To consider these clusters as elements on which the multidimensional scaling are performed often leads to a better configuration. Therefore s(A,B) should be considered as the similarity on which the projection should be performed. If we perform the clustering by the procedure described in the previous subsection, we will be consistent in the whole process of the clustering and the scaling.

Another application of s(A,B) is classification of a new individual based on the model on which s(A,B) is defined. Suppose that we have categories $\{X_1,...,X_K\}$ which was generated by the above procedure or by some other way. An easy way to classify a new individual w is to calculate $s(w,X_p)$, $p=1,...,K$ and allocate w into the category $X_s$ which satisfies

$$s(w,X_s) = \max_{1 \leq i \leq K} s(w,X_p) .$$

## 4. Optimization of dendrograms

### 4.1 A problem in the single linkage and in the complete linkage

The single linkage and the complete linkage methods are the two most well known algorithms among various techniques of the hieararchical agglomerative clustering. According to their applications, some researchers prefer the single linkage; others prefer the complete likage. When we observe the dendrograms produced by these two methods we frequently notice that the dendrogram representations have a problem. In a dendrogram we observe not only the generated groups but also the process of the generation of the clusters one by one. If the merges of the clusters are concentrated at a particular level of similarity, it is difficult to see the structure of the dendrogram. This problem is typical in the single linkage and in the complete linkage, although other algorithms such as the average linkage method also have the same problem. In the single linkage method frequently a major part of the merges occurs at higher levels of similarity in the whole process of the generation of dendrograms. In the complete linkage a considerable part of the merges is inclined to occur at lower levels of similarity. Everitt (1980, p.87) showed several examples of single likage clustering in which we observe that 85% of the merges are occurred in an interval of the higher similarity whose length is 20% of the length of the whole interval of the similarity of the merges. We find that to see clearly structures of the generated clusters are frequently difficult in such a kind of dendrograms. If we denote the level of similarity for each merge as $m_1$, $m_2, \ldots, m_{n-1}$, and if we make a histogram of these data of the levels of the merges, we will obtain a histogram like the one shown in Fig. 2, where a sharp peak of the histogram is observed.

4.2 Histogram flattening

Let us note again that a dendrogram is a tree-like figure with one axis that shows the levels of the merges based on similarity. Every point of branch of the dendrogram can be projected onto the axis: the point of projection shows the level of similarity at which the two clusters are merged.

There are two ways for output of a dendrogram about the treatment of the level of similarity of the merges. In a discrete treatment the axis shows a number of discrete levels of the merges. Let us denote this number as c. If we assume that $m_1 < m_2 <\ldots< m_{n-1}$, a usual way of the discrete treatment gives the classes of intervals $[m_1,z_1],(z_1,z_2],\ldots,(z_{c-1},m_{n-1}]$ of an equal length (i.e., $z_1-m_1=z_2-z_1=\ldots=m_{n-1}-z_{c-1}>0$.) Each level of a merge of two clusters is put into some of these classes. Merges in the j-th class are represented by the j-th discrete level on the axis. In other words, the projection of the merges in the j-th class are at the j-th level on the axis. (See Fig. 3.) This kind of discretization is necessary to print a dendrogram on a usual type of printers such as line printers without a graphic output option.

On the other hand when we have a graphic printer we can use a continuous treatment in which the axis is continuous: a projection of the j-th merge is $f(m_i)$ with a continuous strictly monotone transformation, without any categorization of the merges.

The problem posed in the previous section should be considered in both the discrete and the continuous treatments. We begin by the discrete case.

As was suggested in the previous section, the difficulty of observing a dendrogram comes from a high peak (or high peaks) of the histogram of the merges. Therefore some computer programs of hierarchical clusterings allow a user to specify freely the levels $z_1,\ldots,z_{c-1}$ for the classes, although default value is of course $z_1-m_1=z_2-z_1=...=m_{n-1}-z_{c-1}$ for equal intervals. These consideration suggests an automatic method of selecting levels $z_1,...,z_{c-1}$ so that the resulting information on the dendrograms is maximized.

Let us represent the levels by a vector $z=(z_1,...,z_{c-1})$ and let $(h_1,...,h_c)$ be a histogram of the merges of pairs of clusters. In other words $h_j$ is the number of $m_i$'s satisfying $z_{j-1}<m_i\leq z_j$. Therefore we represent $h_j=h_j(z)$ as functions of the levels. A natural formulation to maximize information is given by an optimization

$$\max \; - \sum_j \frac{h_j(z)}{n-1} \log \frac{h_j(z)}{n-1} \qquad (3)$$

$$\text{subject to } \; z_1 < z_2 <...< z_{c-1} \; .$$

When the problem is formulated as above, this method has the same form as the histogram flattening which is well-known in the field of image processing (Rosenfald, Kak, 1976). It is easy to maximize the above criterion in an approximate way, since the number of the merges is not very large in general. Therefore we omit the detail of the algorithm for the optimization.

### 4.3 Optimization of the dendrogram in the continuous case

A similar but somewhat different method can be considered for the continuous case, where we do not have any discrete class of the merges. An analogous way for the formulation is to define $z_i$'s not as the ends of the intervals of the classes but as the coordinates of the projections of the merges, namely,

$z_i=f(m_i)$, $i=1,\dots,n-1$ $(c=n)$. The simplest choice is that f is an affine transformation, in which case we will obtain the original dendrogram. (See Fig.4.)

If we consider an optimization

$$\max \left| \sum_j (z_j-z_{j-1}) \log (z_j-z_{j-1}) \right| \qquad (4)$$

$$\text{subject to } z_1 < z_2 <\dots< z_{n-1}, \quad z_{n-1}-z_1=\text{const}.$$

it is easy to see that the optimal solution is given by $z_2-z_1=\dots=z_{n-1}-z_{n-2}$. This solution corresponds to the histogram flattening in the previous subsection. Unfortunately the above solution is not useful to a user of the hierarchical clustering, since the output expresses only the order of the merges. A good way to deal with the problem of optimizing dendrograms in the continuous case is to restrict the class of admissible transformations for the criterion (4).

Let us consider a piecewise linear transformation

$$f(x) = \begin{cases} \dfrac{e-z_1}{d-m_1}(x - m_1) + z_1, & z_1\le x<d \\[2ex] \dfrac{z_{n-1}-e}{m_{n-1}-d}(x - d) + e, & d\le x\le m_{n-1} \end{cases}$$

for a fixed $z_1$ and $z_{n-1}$, and PL be the class of all piecewise linear transformations of the above form with all $m_1\le d\le m_{n-1}$ and $z_1\le e\le z_{n-1}$. Then consider

$$\max \quad | \Sigma \ (f(m_i)-f(m_{i-1})) \log (f(m_i)-f(m_{i-1})) | \qquad (5)$$

subject to $f \varepsilon PL$ .

Since the computation of an approximate solution is not difficult, we omit the detail.

Remark This method of restricting admissible transformations to a class of piecewise linear functions is applicable to the discrete case. We studied this method in picture enhancement problem (Miyamoto and others, 1985). The application of this method to optimization of dendrogram is straightforward and we omit the detail.

Remark Another motivation for optimization of dendrograms comes from the desire to compare two dendrograms. Frequently we wish to compare two dendrograms of the same set of entities by different algorithms of the hierarchical clustering to check whether they have similar structures or not. In such a case it is much better to compare those two dendrograms in their optimized forms, in other words, in their enhanced forms.

## 5. Conclusion

In the present paper we dealt with solely hierarchical methods of cluster analysis. Various algorithms of nonhierarchical clustering have been published including those of fuzzy clustering (e.g., Bezdek, 1981). Nevertheless, here we emphasize the significance of hierarchical cluster analysis. Successful application of the cluster analysis can be divided into two types. In one type methods of analysis are less developed. One does not have sufficient prior knowledge, nor experience about the nature of the clusters. In these applications researchers try to increase their knowledge through clustering: they compare a number of different clusters to find what is more appropriate structure to fit their intuition end experiences. For these applications hierarchical cluster analysis is more adequate. In the other type of the successful applications methods of analysis are more developed. Experiences have been accumulated and one knows an approximate number of clusters to be found. For example, application to remote sensing belong to this category. In the latter applications nonhierarchical methods such as ISODATA (Ball and Hall, 1965) are successful.

In this paper it has been implicitly assumed that we are dealing with the former type of applications with little prior knowledge. In these applications sometimes no appropriate framework has been established. Therefore researchers are trying to find what is an adequate tool of analysis. What is important in such a case in general is to provide tools that is easy to apply without much prerequisite, and the hierarchical cluster analysis is one of such tools. Indeed, the hierarchical methods are easy to apply, nevertheless, they have various problems, a part of which has been considered in this paper. The hierarchical methods of cluster analysis can be called as a "small" tool in the sense that they are easy to apply to various real problems. On the other hand, one should not draw a strong conclusion only by the result of the clustering. One should check the result of the clustering with other type of data or knowledge to obtain a clear understanding of the system.

## References

Anderberg, M. R. (1973) Cluster Analysis for Applications, Academic Press, New York.

Ball, G. H. and Hall, D. J. (1965) ISODATA, A novel method of data analysis and pattern classification, AD699616, Stanford Res. Inst., Menlo Park, California.

Bezdek, J. C. (1981) Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York.

Everitt, B. (1980) Cluster Analysis, 2nd ed., Halsted Press, New York.

Hartigan, J. A. (1975) Clustering Algorithms, Wiley, New York.

Kendall, M. G. and Stuart, A. (1973) The Advanced Theory of Statistics, Vol.2, 3rd ed., Griffin, London.

Kruskal, J. B. (1964) Nonmetric multidimensional scaling: a numerical method, Psychometrica, 29, 115-129.

Miyamoto, S., Oi, K., Naito, M., Shimizu, A. (1985) Enhancement of laser radar images by a class of piecewise linear transformations of gray levels based on entropy criteria, Proc. IEEE Workshop on Language For Automation, Palma de Mallorca, Spain, 265-270.

Miyamoto, S., Nakayama, K. (1986) Similarity measures based on a fuzzy set model and application to hierarchical clustering, IEEE Trans., Syst., Man, and Cybern., 16, 3, 479-482.

Miyamoto, S., Oi, K., Abe, O., Katsuya, A., Nakayama, K. (1986) An information retrieval and data analysis system designed for surveyed data of association tests, Proc. 4th IFAC Symp. on Large Scale Systems, Theory and Applications, Zurich, Switzerland, to appear.

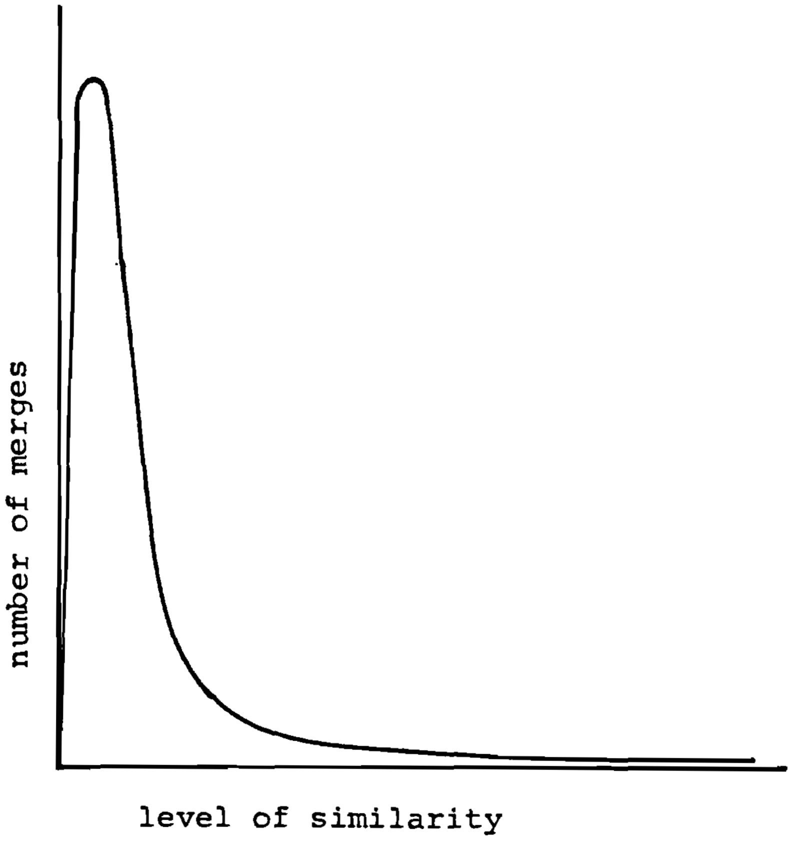
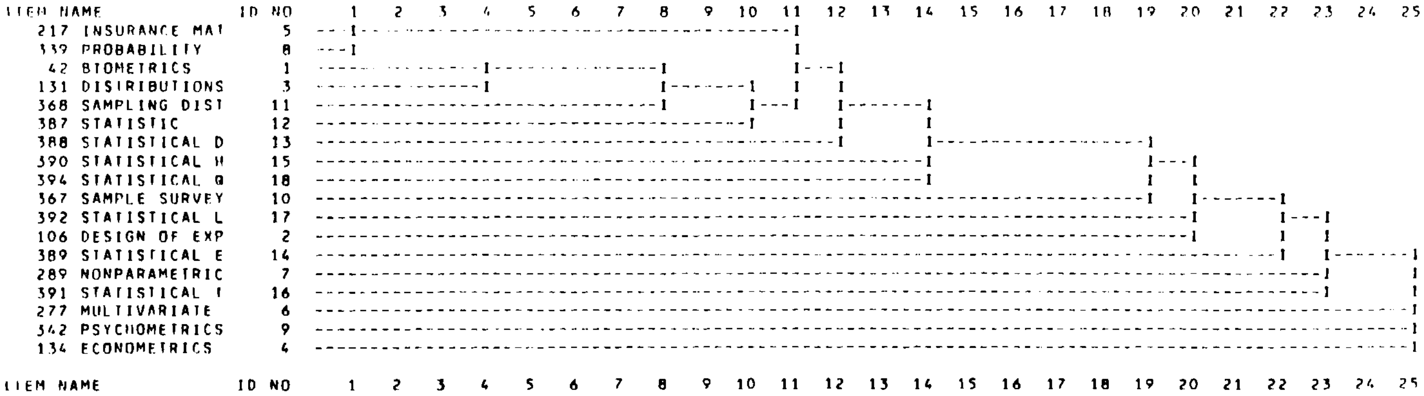


Fig. 2 A typical histogram of merges occurred in the dendrograms by the single linkage method.



```
ITEM NAME           ID NO      1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25
  217 INSURANCE MAT     5   ---I---------------------------------------I
  339 PROBABILITY       8   ---I                                       I
   42 BIOMETRICS        1   ---------------I---------------I           I---I
  131 DISTRIBUTIONS     3   ---------------I               I-------I   I   I
  368 SAMPLING DIST    11   -------------------------------I       I---I   I-------I
  387 STATISTIC        12   ---------------------------------------I       I       I
  388 STATISTICAL D    13   -----------------------------------------------I       I-------------------I
  390 STATISTICAL H    15   -------------------------------------------------------I                   I---I
  394 STATISTICAL Q    18   -------------------------------------------------------I                   I   I
  367 SAMPLE SURVEY    10   ---------------------------------------------------------------------------I   I-------I
  392 STATISTICAL L    17   -------------------------------------------------------------------------------I       I---I
  106 DESIGN OF EXP     2   -------------------------------------------------------------------------------I       I   I
  389 STATISTICAL E    14   ---------------------------------------------------------------------------------------I   I-------I
  289 NONPARAMETRIC     7   -------------------------------------------------------------------------------------------I       I
  391 STATISTICAL I    16   -------------------------------------------------------------------------------------------I       I
  277 MULTIVARIATE      6   ---------------------------------------------------------------------------------------------------I
  342 PSYCHOMETRICS     9   ---------------------------------------------------------------------------------------------------I
  134 ECONOMETRICS      4   ---------------------------------------------------------------------------------------------------I

ITEM NAME           ID NO      1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25
```

Fig.3 An example of a dendrogram with discrete treatment. The numbers 1-25 below the dendrogram show 25 levels on the axis. (This dendrogram was copied from Miyamoto and others, Development of a computer program package for bibliometrics, Report of a research supported by the Grant in Aid for Fundamental Scientific Research of the Educational Ministry in fiscal 1983, in Japanese.)