Recent Developments in Statistical Inference and Data Analysis; K. Matusita, editor

A CRITERION FOR CHOOSING THE NUMBER OF CLUSTERS IN CLUSTER ANALYSIS

KAMEO MATUSITA AND NOBORU OHSUMI

The Institute of Statistical Mathematics
Tokyo, JAPAN

In the recent work, attention in cluster analysis is directed towards the development of various criteria that may be used for evaluating the clustering process, rather than clustering procedures. In practice, it is an important problem to choose the number of clusters or evaluate the clustering process by changing variously the number of clusters. In this paper, we shall propose use of the affinity to solve such a problem.

INTRODUCTION

In cluster analysis, it comes into question to settle on the number of clusters. In fact, various attempts have been made to handle this problem, but the most important point in such a case is how to define the term "cluster". Most proposed definitions consist of statements such that a cluster is a set of observations which are similar to each other. But these definitions are very vague. In any case, we must examine partitioning the given data set into various sets of clusters. However, it is impossible to assess all possible ways of partitioning.

The clustering procedures which we treat in this paper are iterative partitioning techniques and based upon the following well-known relationship :

$$T = W + B \qquad (1)$$

where T is the total dispersion matrix, W the within-cluster dispersion matrix, that is, $W = \sum_{i=1}^{k} W_i$ where W_i is the dispersion matrix for the i-th cluster C_i, and B the between-clusters dispersion matrix when the observations are partitioned into a set of k clusters. These techniques, as many others, attempt to minimize tr(W), where W is given in expression (1). In view of this, we assume that the clusters we seek are spherical and relatively compact in shape. Accordingly, the problem is how to choose a reasonable number of clusters of such shape, and to approach the above objective as far as possible by iterative procedure. In general, the solution obtained by these procedures is local optima. Moreover, there has been no way of knowing whether or not the best optimal solution has been reached. To tackle this problem we shall propose the affinity as a criterion to examine a set of clusters formed by the iterative partitioning procedure.

EVALUATION BASED ON THE ADJUSTED AFFINITY OF CLUSTERS

In the previous papers [1], [2], we have discussed application of the affinity to the clustering process. In particular, we are trying to make improvements of algorithm and make a program-package for computers. Several new ideas are included in the present paper, for example, that of the function which controls the cluster

size and the within-variance in each cluster, the treatment for the occurrence of singularity, and so on. For making clear the idea of the functions, we shall illustrate the experimental results of typical partitioning procedures, the k-means method and the ISODATA method. The optimization in the k-means method and the ISODATA method is performed by minimizing tr(W). That is, this procedure is identical with minimization of the sum of squared euclidian distances between individuals in a cluster. However, both methods are essentially different in the following points.

First, in the k-means method the number of clusters is initially fixed and the minimization of tr(W) is achieved by replacement and exchange of individuals. On the other hand, the procedure of the ISODATA method not only minimizes tr(W), but also utilizes the following characteristics concerning the algorithm :

1) to delete temporarily outliers or clusters consisting of small numbers of individuals,
2) to carry out repeatedly the lumping (or merging) process and the splitting process of clusters, and to perform the minimization of tr(W) by changing the number of clusters,
3) to choose between splitting and lumping locally by evaluating the variation of each cluster.

After all, the ISODATA method differs basically from the k-means method in the point of the automatical division or combination. Besides, we have improved each point described above in our computer programs. More detailed descriptions are presented in the references [3], [4], and so on.

For a set of clusters $\{C_1, C_2, \ldots, C_k\}$ we represent the degree of separation of the set of clusters by the affinity

$$\rho_k (C_1, C_2, \ldots, C_k) = \frac{\prod_{i=1}^{k} | U_i |^{1/2k}}{| 1/k \, \Sigma U_i |^{1/2}} e^{Q/2k} \qquad (2)$$

where

$$Q = (\Sigma U_i \bar{x}_i)' (\Sigma U_i)^{-1} (\Sigma U_i \bar{x}_i) - \Sigma \bar{x}_i' U_i \bar{x}_i \quad ,$$

$\bar{x}_i$ being the mean vector of the observations in cluster C_i, and U_i the inverse of the variance-covariance matrix of C_i. Further, to make more comprehensive the behavior of affinity ρ_k over changing k, we take $\rho_k^* = (1/k) \log \rho_k$ as *average compactness* or *spread per cluster*. We call ρ_k^* the *adjusted affinity*.

By the way, if we want to take into account the criterion of expression (2), it will be more reasonable to utilize the generalized distance rather than the squared euclidian distance, that is $D_G = (x_j - \bar{x}_i)' U_i (x_j - \bar{x}_i)$, where x_j is the observed vector of the j-th individual in the i-th cluster and $\bar{x}_i$ is the mean vector of the i-th cluster.

However, when utilizing the generalized distance, the result of clustering is strongly influenced by the shape of data, (i.e., by the size and the direction of variation of each group generated by clustering, such as the phenomenon observed frequently in discriminant analysis). Thus, in the case where there are several elongated clusters and the variance-covariance matrix in each cluster is relatively large, it will be recommended to use the criterion of the sum of squares and to adopt a little more large number of groups than that of those which will seem to exist really.

To illustrate our procedure, we construct four sets of data in the following way. First, let $G_1, G_2, \ldots, G_5$ be three-dimensional Gaussian distributions with mean

vectors

$$\mu_1 = (-3.0, -2.0, 2.0) \qquad \mu_2 = (0.0, -2.0, -2.0)$$
$$\mu_3 = (0.0, 2.0, 2.0) \qquad \mu_4 = (-0.5, 2.0, -2.0)$$
$$\mu_5 = (0.0, 0.0, 0.0)$$

and variance-covariance matrices

$$V_1 = \begin{bmatrix} 1.00 & -0.42 & -0.10 \\ & 0.36 & 0.42 \\ & & 1.00 \end{bmatrix},$$

$$V_2 = V_3 = V_4 = V_5 = \begin{bmatrix} 1.00 & 0.45 & 0.60 \\ & 0.36 & 0.15 \\ & & 1.00 \end{bmatrix},$$

respectively. We took

100 observations (random sample of size 100) from G_1 and G_4, respectively,
200 observations from G_2,
300 observations from G_3 and G_5,

respectively, and made the mixture of these 1000 observations. Denote it by A. Similarly, we formed two more sets of the same structure, B, C. Further, we formed a set of 2500 observations in a similar manner. Denote it by D. To these sets A, B, C, D we applied the two methods mentioned above.

First, we examine the results by the k-means method. A part of the experimental results is shown in Table 1. It is seen that the indicator ρ_k^* attains values near its minimum when the number of clusters is 5 to 8 in the range 2 to 10. This is commonly observed at each set. For example, let us look at the case A shown in Table 1, and compare the given data set A with the clusters actually obtained. The set of data in Figure 1 actually consists of several relatively well separated spherical groups, but it seems difficult to recover the groups as given initially. In fact, ρ_k^* takes a lower value when the data is partitioned into seven or eight clusters, and we cannot exactly detect the five groups.

However, it can be seen that the number of clusters chosen by the value judgement of ρ_k^* is reasonable. In fact, we can verify the validity of the judgement by the visual observation of the scatter diagrams for the sets of data classified into five or eight clusters as shown in Figures 2 and 3. The behavior of ρ_k^* represents clearly the well-known feature that the clustering criterion such as minimization of tr(W) pertains to constructing a spherical and well-condensed group structure on the data. Besides, the result of eight clusters, (see Figure 3), is preferable to that of five clusters. Because we can observe gaps or moats for clusters C_1 and C_3 in Figure 2, which are known as "wild-shot".

On the other hand, it seems that the clusters of the eight groups are like balls in shape, respectively, but we can adequately grasp the traits of data by linking together clusters. As is seen in this example, it will be reasonable to form a little more groups.

Secondly, we applied the ISODATA method to the sets of data A, B, C and D. In the ISODATA method, the number of clusters can be altered variously in some range without the number of groups previously specified as in the k-means method. In this case, when the number of iteration of the clustering process was ten for each set of data, we obtained five clusters as an optimum set of clusters. Nevertheless, we can see almost the same result as that of the k-means method. The values of ρ_k^* are as follows :

$\rho_k^* = -0.69085$ (for the case A) (n=1000)

$\rho_k^* = -1.18481$ (for the case B) (n=1000)
$\rho_k^* = -1.06616$ (for the case C) (n=1000)
$\rho_k^* = -0.76247$ (for the case D) (n=2500)

In the case A, it can be observed that the result of the ISODATA procedure agrees with that of the k-means method, but, of course, it happened by accident. Especially, in the case D, forming the cross-classified membership table between the actual five groups and the five clusters obtained by the ISODATA procedure, we can obtain Table 2. This table illustrates clearly that ρ_k^* is effective as a criterion for evaluating the clustering process or choosing the number of clusters.

CONCLUSION

In general, it is obvious that clustering procedures depend strongly upon the algorithm and the criterion that generates the set of clusters. However, as is described above, investigation of the behavior of ρ_k^* indicates with objectivity that each cluster formed is likely to be nearly spherical in shape. A clustering procedure may be evaluated by tracing the value of ρ_k^*. Though it is impossible to check all possible partitions, we can search approximately for a reasonable partition into clusters by using a feasible method and make comparisons between several results of clustering.

Finally, we add that a clustering program-package for computers, called MINTS (MINTS is an abbreviation of "MINi Numerical Taxonomy System"), has been prepared to carry out the two procedures proposed here.

Table 1.

Values of adjusted affinity ρ_k^* obtained by the k-means method

Iteration	Number of clusters	Adjusted affinity ρ_k^*			
		Case A	Case B	Case C	Case D
2	2	-0.271	-0.225	-0.239	-0.248
3	3	-0.639	-0.911	-0.413	-0.745
4	4	-0.549	-0.848	-0.823	-0.628
5	5	-0.690	-1.176	-1.066	-0.744
6	6	-1.125	-1.221	-1.085	-0.836
7	7	-1.149	-1.260	-1.324	-1.138
8	8	-1.240	-1.082	-1.165	-1.049
9	9	-0.937	-0.997	-1.060	-0.970
10	10	-0.864	-0.926	-0.995	-0.863

Table 2.

Cross-classified table obtained by the ISODATA method

Clusters	Actual groups G_5	G_1	G_2	G_3	G_4	Cluster size	Obtained mean vector
C_1	612		42	59	2	715	(0.01 0.02 0.08)
C_2	5	247		3		255	(-2.98 -1.96 1.95)
C_3	59	3	458			520	(-0.35 -1.96 -2.10)
C_4	49			672	8	729	(0.28 1.99 2.12)
C_5	25			16	240	281	(-0.56 1.84 -1.93)
Size of group	750	250	500	750	250	2500	

REFERENCES

[1] Matusita, K., Cluster Analysis and Affinity of Distributions, *Recent developments in statistics* (eds.) J. R. Barra et al., (North-Holland Publ. Co. 1977).

[2] Matusita, K. and Ohsumi, N., Evaluation Procedure of some Clustering Techniques, Paper presented at the Franco-Nippono Seminar (1978).

[3] Hall, D. J., Khanna, D. The ISODATA method, Computation for the Relative Perception of Similarities and Differences in Complex and Real Data, *Statistical Methods for Digital Computers,* Volume III of *Mathematical Methods for Digital Computers,* (eds. Enslein, Ralston and Wilf), (John Wiley & Sons, 1977).

[4] Tou, J. T., Gonzalez, R. C. *Pattern Recognition Principles,* (Addison-Wesley, 1974).

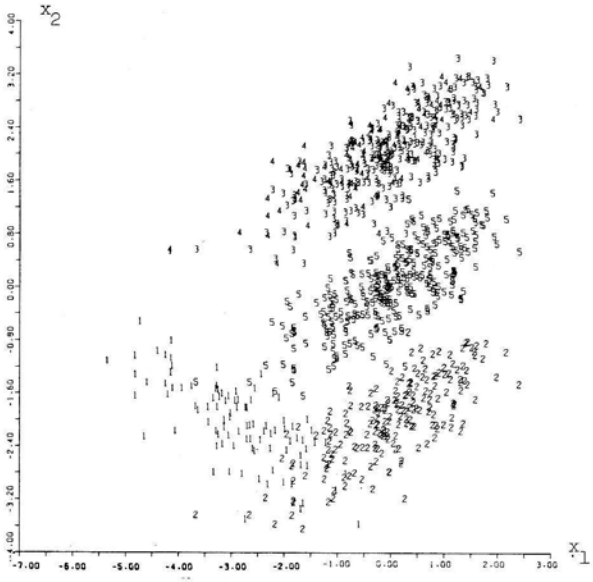

Figure 1-(a).

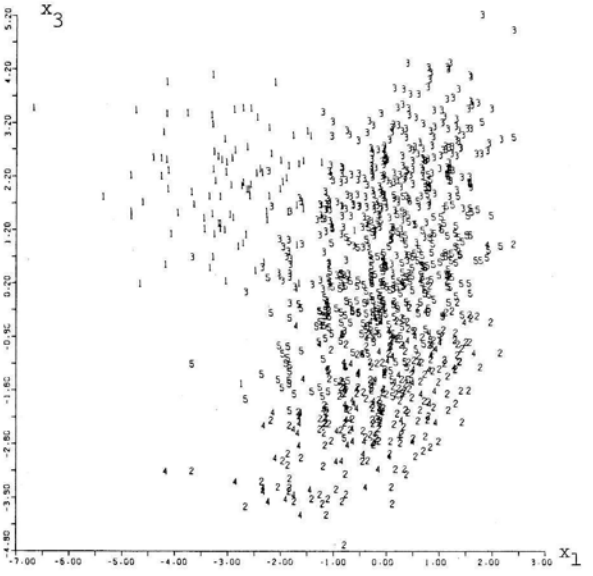

Figure 1-(b).

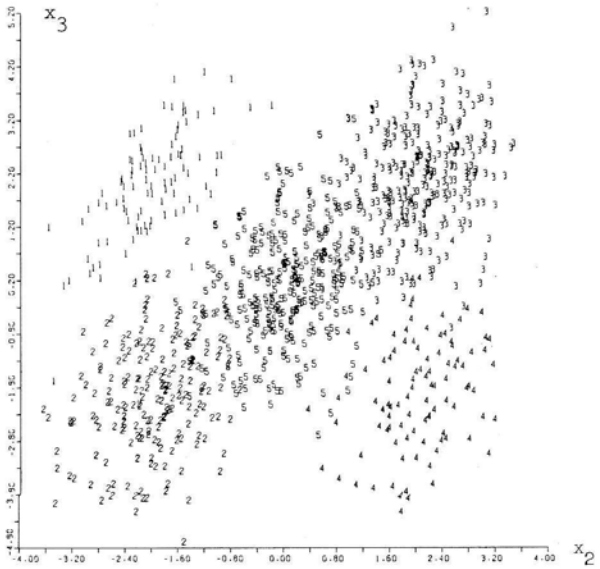

Figure 1-(c).

Figure 1. Five groups generated from the mixture of three-dimensional Gaussian distributions (in the case of data set A).

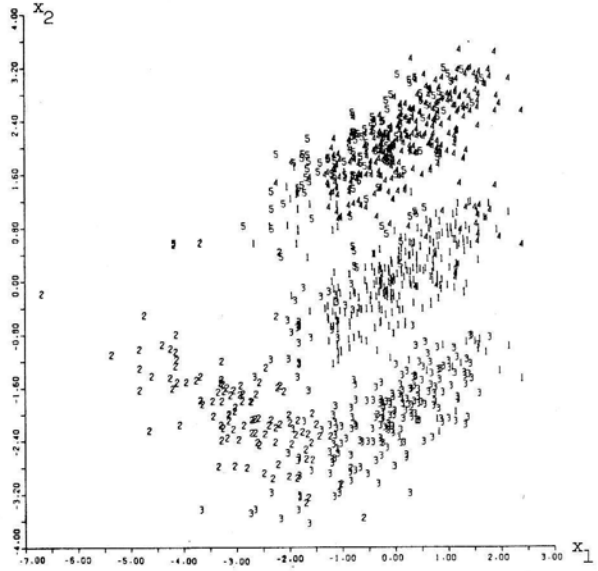

Figure 2-(a).

In this figure, we can slightly observe the presence of moats.

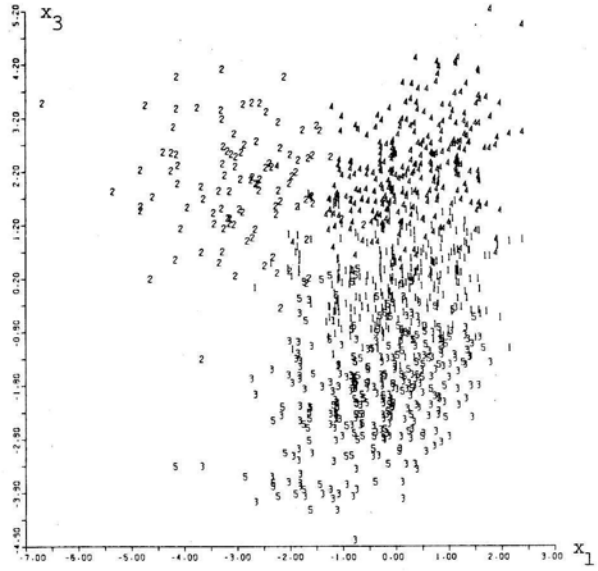

Figure 2-(b).

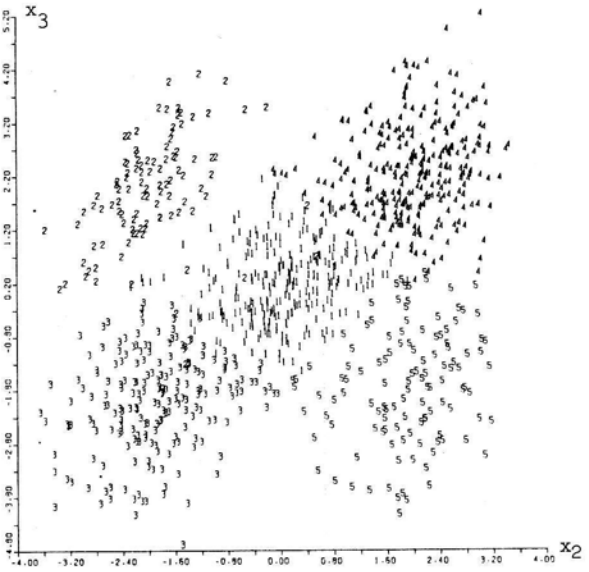

Figure 2-(c).

Figure 2. The set of data partitioned into five clusters by k-means method in the case of A shown in Table 1 ($\rho_K^* = -0.691$).

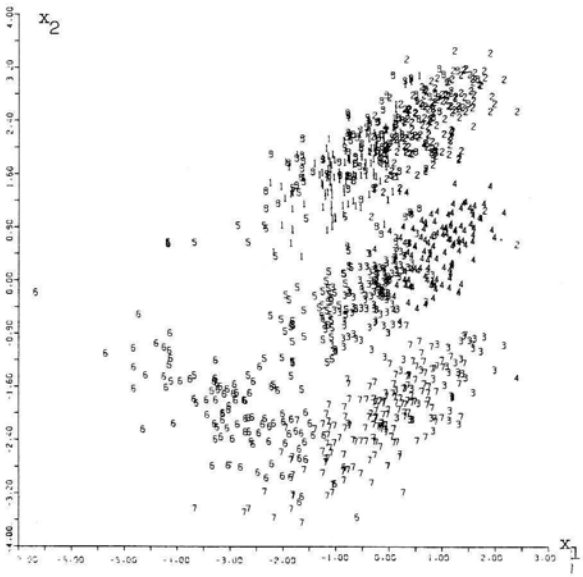

Figure 3-(a).

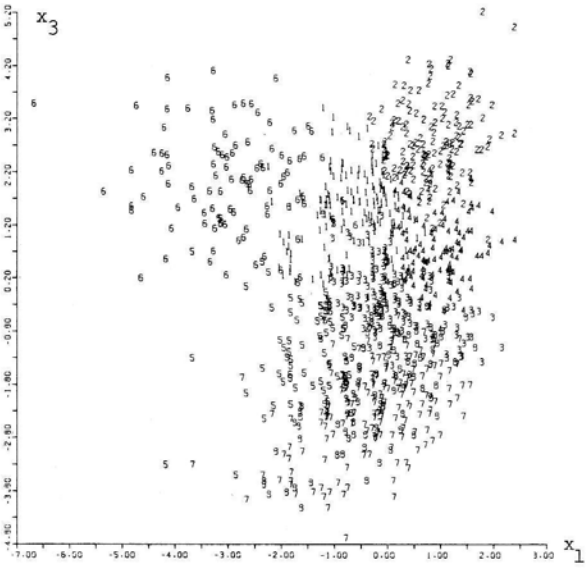

Figure 3-(b).

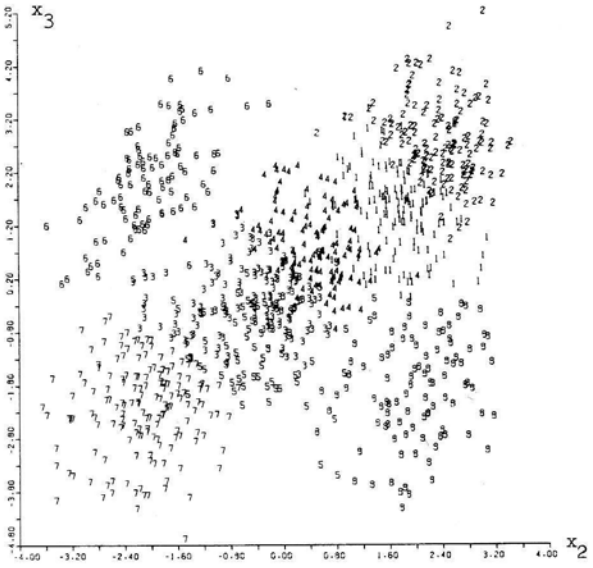

Figure 3-(c).

Figure 3. The set of data partitioned into eight clusters by k-means method in the case of A shown in Table 1 ($\rho_k^* = -1.240$).

REPRINT FROM:

RECENT DEVELOPMENTS IN STATISTICAL INFERENCE AND DATA ANALYSIS

Proceedings of the International Conference in Statistics in Tokyo

Edited by

K. MATUSITA
The Institute of Statistical Mathematics
Tokyo

1980

NORTH-HOLLAND PUBLISHING COMPANY
AMSTERDAM • NEW YORK • OXFORD