Program CLAN: A general-purpose, user-oriented cluster analysis program

T. D. KLASTORIN

*Department of Management and Organization DJ-10
University of Washington, Seattle, Washington 98195*

and

ROBERT LEDINGHAM

*Department of Health Services SC-37
University of Washington, Seattle, Washington 98195*

Program CLAN is a general-purpose, user-oriented cluster analysis program that includes a number of procedures for searching for homogeneous groups of taxonomic units. Program CLAN allows a user to search for two types of classification structures: (1) hierarchical structures (when groups are nested into other groups) and (2) partitions (when taxonomic units are classified into mutually exclusive and exhaustive groups).

The program consists of three parts. In the first part, the weighted characteristic vectors that describe the taxonomic units are combined into scalar values that express a degree of similarity between all pairs of taxonomic units. In this section, numerous options are available to the user, including the ability to standardize the raw data, compute principal components, automatically employ regression analysis to estimate variable weights, and compute any number of similarity measures.

The second part of the CLAN program uses the matrix of pairwise similarity measures to search for homogeneous partitions of taxonomic units. If a hierarchical method is specified, the program finds a hierarchy of groups called a dendrogram (illustrated in Figure 1) that is subsequently evaluated to identify possible partitions. If an iterative method is specified, the taxonomic units are partitioned directly into the required number of groups.

In the present version, there are seven hierarchical agglomerative cluster analysis algorithms in the program and one iterative clustering algorithm. The seven hierarchical algorithms include (1) the single linkage method (Sneath & Sokal, 1973), or minimum method (Johnson, 1967), (2) the average linkage between groups method (Sneath & Sokal, 1973), (3) the complete linkage method (Lance & Williams, 1967), or maximum method (Johnson, 1967), (4) Ward's (1963) suboptimization method, (5) Gower's (1967) median method, (6) the average linkage within merged groups method (Sneath & Sokal, 1973), and (7) the centroid method (Gower, 1967). The iterative algorithm is MacQueen's (1967) k-means

The development of this program was supported in part by HEW Contract 600-76-0143.

Figure 1. Example dendrogram for six taxonomic units.

algorithm (see also Hartigan & Wong, 1979). A complete description of these cluster analysis algorithms is given by Anderberg (1973).

The third part of the program allows a user to statistically test and compare the various partitions that may have been found. Included in this section are statistics for testing consistency among partitions (i.e., how much do various partitions differ with respect to taxonomic units similarly grouped?). A method for testing whether the groups in one partition merely represent finer subdivisions of groups in another partition is also available (Klastorin, in press).

The CLAN program contains a number of novel features not found in other cluster analysis programs; perhaps the most notable difference is the program's method for analyzing dendrograms created by a hierarchical algorithm. To illustrate the process, consider the dendrogram in Figure 1. In this case, each horizontal line segment in the dendrogram identifies a unique group; it is assumed that the length of a group's respective line measures the distinctiveness of that group (i.e., the longer the horizontal line, the more distinct the group). Given this assumption, a measure of distinctiveness (called expected distinctiveness) can be defined for each group identified in a dendrogram. By summing these measures for all groups in a partition, a distinctiveness measure is defined for any given partition; expected distinctiveness values for each partition defined by a vertical line through each joining class level (the number of joining class levels is a parameter set by the

user in this program) are given in Figure 1. However, since the partition that maximizes expected distinctiveness may not be a partition defined by a vertical line drawn through one of the joining class levels, the program contains an algorithm that searches each dendrogram for the most distinct partition. Details of this algorithm are explained in Klastorin (Note 1).

Other unique features in the CLAN program include the ability to incorporate constraints in the partition search process (e.g., a user might wish to specify an upper bound on group sizes) and a number of partition assessment statistics for testing whether the identified partitions differ from randomly determined partitions (Klastorin, Note 2).

Input. A user can input raw data directly or input similarity measures previously computed. Variable weights can be input directly or calculated by the program if a dependent variable is specified. Labels for the taxonomic units can also be read in if the user wishes. All input is flexible; the user specifies whether cards or disk files are to be used. The format of the control statements is similar to those used by the SPSS program (Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975).

Output. The user can print out the raw data, transformed data, similarity measures, and so on, if desired. If a hierarchical algorithm is used, a dendrogram similar to the one in Figure 1 is printed, as well as complete descriptive information on each partition found. If requested, assessment statistics and significance levels are printed for each partition found. Files containing intermediate results, such as the similarity matrix and partitions, can be printed or saved on disk storage.

Computer Specifications. The program is written in FORTRAN IV and has been extensively run on the University of Washington CDC 6400/Cyber 73 system using both the MNF and FTN compilers. The program can analyze up to 350 taxonomic units in 260K₈ words of memory; the memory requirements can be significantly reduced if fewer taxonomic units are analyzed (e.g., 100 units can be analyzed with 100K₈ words). The minimum number of words required by this pro-

gram is 60K₈. The program uses no overlays and should be easily transferable between systems.

Availability. User documentation and program information are available at no cost from T. D. Klastorin, 126 Mackenzie Hall DJ-10, University of Washington, Seattle, Washington 98195. A program tape is available at a nominal cost (covering the cost of the tape and mailing).

REFERENCE NOTES

1. Klastorin, T. D. *An objective method for evaluating hierarchical cluster analysis* (Working Paper). Seattle: University of Washington, 1979.
2. Klastorin, T. D. *On assessing cluster analysis results* (Working Paper). Seattle: University of Washington, Graduate School of Business, 1979.

REFERENCES

- ANDERBERG, M. R. *Cluster analysis for applications*. New York: Academic Press, 1973.
- GOWER, J. C. A comparison of some methods of cluster analysis. *Biometrics*, 1967, **23**, 624-637.
- HARTIGAN, J. A., & WONG, M. A. A k-means clustering algorithm. *Journal of the Royal Statistical Society*, 1979 (Series C), **28**, 100-108.
- JOHNSON, S. C. Hierarchical clustering schemes. *Psychometrika*, 1967, **32**, 241-254.
- KLASTORIN, T. D. Merging groups to maximize partition comparison. *Psychometrika*, in press.
- LANCE, G. N., & WILLIAMS, W. T. A general theory of classificatory sorting strategies, 1. Hierarchical systems. *Computer Journal*, 1967, **9**, 373-380.
- MACQUEEN, J. B. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, June 21-July 18, 1965* (Vol. 1). Berkeley: University of California Press, 1967.
- NIE, N. H., HULL, C. H., JENKINS, J. G., STEINBRENNER, K., & BENT, D. H. *Statistical package for the social sciences* (2nd ed.). New York: McGraw-Hill, 1975.
- SNEATH, P. H., & SOKAL, R. R. *Numerical taxonomy*. San Francisco: Freeman, 1973.
- WARD, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 1963, **58**, 301.

(Accepted for publication June 26, 1980.)