Historia Mathematica Vol. 11, No. 1, 1998, 52-57

A Historical Study on Statistical Packages in Cluster Analysis

서경대학교 응용통계학과 **이승우**

Abstract

Since cluster analysis encompasses many diverse techniques for discovering structure within complex bodies of data, it has been employed as an effective tool in scientific inquiry. Recent works on cluster analysis softwares carried out by SAS, SPSS, S-PLUS and BMDP are briefly summarized and investigated in this paper.

The inferred statistical package for windows executing a way for data analysis in modern statistical techniques has several merits superior to other packages. Especially, S-PLUS can be designed and tried out much faster than other statistical packages. S-PLUS provides a graphic which is interactive, informative, flexible ways of looking at data. Also, if a statistical computation time is long and programs are complex, these can be shorten by providing interfaces to the UNIX systems (or C, Fortran).

0. Introduction

The aim of cluster analysis is to divide a set of objects into constituent groups called classes, clumps and clusters. According to a chosen criterion, the members of any one group by using cluster analysis techniques differ from one another as little as possible. The objects are each designated by a set of values of variables of different kinds generally. The objective is to enable one to acknowledge and to construe in a plausible way, an existing structure of such a set of objects, to partition the objects and thus to find a data reduction, or to extract a hypothesis or basis for the prophecies of coming events in the future. Cluster analysis can be of use in almost all of the empirical sciences. Indeed, very diverse procedures have been used successfully, for varying periods, in such fields as psychology, anthropology, medicine, criminology, biology,

geology and archaeology as well as in the social sciences, engineering, computer science, and the economic sciences (especially market research).

Clustering techniques seek to divide a set of data into groups or clusters. Ideal data for such a statistical analysis would yield clusters so obvious that they could be picked out, at least in small-scale cases, without the need for complicated mathematical techniques and without a precise definition of the term 'cluster'. In two or three dimensions, for instance, we could examine the data visually, and so identify any clusters present. However, things are so complicated in practice and consequently there has been a great flourish of clustering techniques, especially during the last decade.

Until recently, the range of clustering options contained in most statistical packages has been severely limited. Any generalization about cluster analysis must be vague because a vast number of clustering methods have been developed in several different fields, with different definitions of clusters and similarity among objects. The variety of clustering techniques is reflected by the variety of terms used for cluster analysis. Cluster analysis techniques may themselves be 'classified' into types roughly;

- Hierarchical techniques; in which the classes themselves are classified into groups, the process being repeated at different levels to form a tree.
- Optimization-partitioning techniques; in which the clusters are formed by optimization of a 'clustering criterion'. The classes are mutually exclusive, thus forming a partition or the set of entities.
- Density or mode-seeking techniques; in which clusters are formed by searching for regions containing a relatively dense concentration of entities.
- Clumping techniques; in which the classes or clumps can overlap.
- Others; methods which do not fall clearly into any of the four previous groups.

These types are not necessarily mutually exclusive and several clustering techniques could be placed in more than one category.

1. Statistical Packages Containing Clustering Software

This section provides a summary of available software for cluster analyses. The variety and amount of cluster analysis software have been surprisingly large for a statistical method with effectively only a twenty-five year history. The statisticians propose continually new methods and clustering software is expanded in the process of

renovation. Many researchers have written and evolved probably hundreds of software packages and popular programs available to perform cluster analysis. Clustering software can be placed into four major categories; (1) collections of subroutines and algorithms, (2) general statistical packages which contain clustering methods, (3) cluster analysis packages, (4) simple programs which perform one type of clustering

The popular packages of statistical programs such as BMDP (Dixon, 1981), SAS (SAS Institute, 1985), SPSS (SPSS, 1986) and S-PLUS (Statistical Sciences, 1993) make an offer the most expedient usage for a research worker using cluster analysis. The character of these packages provides nonprogrammers with relatively easy access to sophisticated statistical methods for a wide variety of research problems. These packages also contain a full range of data screening and manipulation methods which help to make complex analyses simple and feasible.

Cluster analysis packages appear the ultimate in flexibility and provide the sophisticated and serious users for convenience. These packages combine the advantages of general statistical packages; (1) integrated control language and data screening and manipulation procedures, with features of interest to user of cluster analysis, (2) a diversity of clustering methods, special diagnostic features and enhanced graphics. Of the greatest importance is that many of the packages contain hard to find clustering methods of analytical procedures which are appropriate for special programs.

Earlier version of the statistical package, SAS, had one clustering method only complete linkage. Since 1985, a recent release of statistical package, SAS, had contained as follows; single linkage, complete linkage and average linkage plus eight other hierarchical agglomerative methods (centroid, density, flexible-beta, McQuitty's similarity analysis, median, two-stage density linkage, Ward's minimum-variance, maximum-likelihood method). The diagnostics of the package in procedure have been expanded since the output provides many information about the solutions.

SPSS had no clustering methods before 1980. Since 1986, SPSS has contained clustering procedures; CLUSTER and QUICK CLUSTER. CLUSTER uses hierarchical agglomerative methods including seven of the most commonly used techniques (single linkage, complete linkage, average linkage, Ward's method, etc.). QUICK CLUSTER uses a k-means method with limited options for starting partitions. Interesting aspects of this procedure are provisions for missing data and the ability to handle large data sets.

BMDP has contained four procedures in order to do clustering analysis; (1) a collection of single, complete and average linkage to cluster variables, (2) an average linkage (centroid sorting) method to cluster cases, (3) a block clustering method to simultaneously cluster cases and variables, (4) an iterative k-means method which forms

partitions among the cases.

2. Doing Cluster Analysis in S-PLUS

The S language developed at AT&T Bell Laboratories by Richard Becker, John Chambers, and Allan Wilks has appeared since 1984. Since S is a language and an interactive programming environment for data analysis and graphic in the statistics, the S language is a very high level language for performing computations. S encourages to compute, look at data, and program interactively, with quick feedback to enable to learn and understand. The primary goal of the S environment is to utilize and encourage good data analysis. That is, by organizing, storing, and retrieving all sorts of data, S provides numerical methods and other computational techniques in order to understand and use data. Since we can write functions in the S language itself, S is useful programming. These functions can build on the power and simplicity of the S language.

S-PLUS is an enhanced, supported superset of the S language. Hence, S-PLUS is a powerful tool for data analysis. S-PLUS for windows, 1993 is a rich graphical data analysis system object oriented programming language and has presentation graphics, exploratory data-analytic methods, modern statistical methods for developing new statistical tools, and extensibility. S-PLUS has several merits superior to other packages. Since S-PLUS is highly interactive, new function can be designed and tried out much faster than other statistical packages. S-PLUS provides a graphic which is interactive, informative, flexible ways of looking at data. S-PLUS is an interpreted language, in which individual language expressions are read and then immediately executed. The great advantage of interpreted languages is that they allow incremental development.

This chapter describes the S-PLUS cluster analysis functions offered by Version 3.2, 1993. Datas that are suitable for these functions are multivariate measurements on objects, and matrices of distances or dissimilarities between objects. The important functions described are as follows; clorder(using to re-order leaves of a cluster tree), cutree(using to create groups from Hierarchical agglomerative clustering), dist(using distance matrix calculation), hclust(using Hierarchical clustering; three heuristic criteria), kmeans(using iterative relocation; sum of squares criterion only, Hartigan's K-Means clustering), labclust(using to label the leaves of a classification tree), mclass(using classification produced by mclust), mclust(using model-based and Heuristic Hierarchical agglomerative clustering; determination of the number of clusters, rubust clustering), mreloc(using model-based iterative relocation for mclust/'mclass'), plclust(using to plot

trees from Hierarchical clustering), subtree(using to extract part of a cluster tree).

Especially, if programs are complex and computing time is long, S-PLUS provides interfaces to other kinds of computing, such as to commands from the UNIX systems or to C or Fortran routines. Thus, by connecting S language and C, a statistical computation time can be reasonably shorten.

3. Applications

C and Fortran are compiled language, since a compiler translate a complete program in the language into the appropriate machine language. Compiled code runs faster and requires less memory than interpreted code. So interfaces to C and Fortran allow to combine the speed and efficiency of compiled code with the robust, flexible programming environment of S-PLUS. Merit of compiled C and Fortran code runs faster than interpreted S-PLUS code. The disadvantage of compiled C and Fortran code can not be flexible and resilient. Mismatching data types and overrunning arrays are just two types of errors that can occur in compiled code but do not occur in S-PLUS code.

The good time to use compiled code is as follows; (1) in cases where we have such code already written and tested, (2) in cases where we can not use S-PLUS vectorized functions to solve our problem without explicit loops or recursions, (3) in cases where we must do a trivial calculation many times.

Compiled code deals only with data types fixed when the code is compiled. The S-PLUS function .C and .Fortran pass only the most basic modes, the numeric ones and character data. If our code does something numerical, it may be fine to convert all the inputs to double precision and return double precision results. If our code rearranges data, however, we probably don't want to change the modes of the data so S-PLUS code would be better than compiled code. The C and Fortran interfaces ignore the class of datasets so they are not object oriented. Both interfaces allow to communicate with the compiled code by passing pointers to vectors of numbers or character strings. It is harder to develop and debug compiled code than S-PLUS functions. Compiled code is not as portable as S-PLUS code. A good strategy is to do as much as possible in S-PLUS code, including error checking, data rearrangements, selections and conversions, storage allocation, and input/output, and use compiled code to do only the numerical or character string calculations required.

S-PLUS runs on many different computers, from supermicros to large mainframe machines. S-PLUS is a software system that runs under the UNIX operating system on

a variety of hardware configurations. S-PLUS grew out of the interests and needs of people doing statistics research, and it is used extensively by statisticians. A wide range of people are presently using S for analytical computing, graphics, and data analysis in diverse areas of financial analysis, statistics research, management and academia.

References

1. Abdelmonem. A. Afifi and Virginia Clark, *Computer-Aided Multivariate Analysis*, Chapman & Hall, 1990.
2. Helmuth Spath, *Cluster Analysis Algorithms,* John Wiley & Sons, 1980.
3. Kanti V. Mardia, John T. Kent and John M. Bibby, *Multivariate Analysis*, Academic Press, 1979.
4. Richard A. Becker, John M. Chambers and Allan R. Wilks, *The New S Language*, AT&T Bell Laboratories, 1988.
5. Richard A. Johnson and Dean W. Wichern, *Applied Multivariate Statistical Analysis*, Prentice Hall, 1992.
6. *SAS/STAT User's Guide, Version 6, Fourth Edition, Vol 1, 2*, SAS Institute, Inc., Cary, NC, 1990.
7. *S-PLUS for Windows User's Manual Vol 1, 2,* Statistical Science, Inc., Seattle, Washington, 1993.