Computer programs for monothetic classification ("Association analysis")

By G. N. Lance and W. T. Williams*

Association analysis is a monothetic method of efficiently subdividing a population of individuals specified by binary attributes. The general *rationale* of such programs is considered, with particular reference to the problems still to be solved and to the computer logic best suited to the fast analysis of large populations. Five existing programs are briefly compared with these considerations in mind.

Introduction

Hierarchical methods of classifying elements into sets are conventionally divided into *divisive* methods, which progressively subdivide the population, and *agglomerative* methods, which progressively fuse the elements; but the distinction is not absolute, and many cluster-seeking methods cannot easily be so specified. A more fundamental distinction lies in the source of the functions which underlie the strategy. In all agglomerative, and some divisive, methods, these are initially calculated between individual elements; the early stages of the analysis are therefore conducted at the lowest available level of information. If, as in true "similarity" methods, fusions once made are irrevocable, eccentricities and errors in the data may direct the subsequent analysis along an unprofitable path. However, in those cases—notably ecology, criminology and sociology—in which monothetic classification by attributes is acceptable or even desirable, a more powerful divisive system is possible, in that the function used for selection of attributes can be calculated over the entire population. The first suggestion of this type known to us is due to Goodall (1953); but the parameter used was inefficient, and the first program using a parameter of high efficiency was that which we wrote for the Ferranti Pegasus (Williams and Lance, 1958); the method was first used in ecology under the name of "association analysis" (Williams and Lambert, 1959, 1960). Similar programs now exist for the Elliott 803 (Exeter)—this program will also run on the Elliott 503 at approximately a hundredfold increase in speed; for the IBM 704 (Pretoria); the English Electric KDF9 (Sydney); and the Control Data 3600 (Canberra). Programs—believed to be copies of the original Pegasus program—are being written both in FORTRAN and ALGOL for the Ferranti Atlas, but specifications are not yet available. The purpose of this communication is to consider the *rationale* of such programs, the problems yet to be solved, and the alternative computer logics available. All existing programs handle qualitative (i.e. binary) data only, although Dale (1964) has suggested a method of extending the system to quantitative data.

1. Statistical considerations

Division parameter

At each stage of the analysis the population under study is to be divided with respect to a single attribute, so that, in the two resulting sub-populations, this attribute is possessed by all members of one and lacked by all members of the other. The variance of the division-attribute is henceforth zero in all sub-populations, and the attribute selected for a division is to be that which is expected to reduce the variance of all other attributes to the greatest possible extent within the resulting sub-populations. Stein (*in litt.*) has pointed out that a linear regression model will provide a simple solution. For consider a set of variates $x_1, x_2, \ldots x_j, x_k \ldots$ standardized so that $Ex_j = 0$ and $Ex_j^2 = 1$. It is required to find that variate which accounts for as much as possible of the total variance. Now, x_j accounts for $(Ex_jx_k)^2/Ex_j^2$ out of the variance of x_k; and therefore the residual mean square of x_k about the best linear prediction of x_k given x_j is

$$Ex_k^2 - (Ex_jx_k)^2/Ex_j^2.$$

Thus the reduction of total variance achieved by x_j is

$$\sum_{k\neq j} \frac{E(x_jx_k)^2}{Ex_j^2}.$$

But since we are assuming that $Ex_k^2 = 1$, the quantity $E(x_jx_k)^2/Ex_j^2$ becomes equal to the squared correlation coefficient; the reduction of variance due to x_j is thus $\sum_{k\neq j} r_{jk}^2$, which in the case of the 2 × 2 table is equivalent to $\sum_{k\neq j} \chi_{jk}^2/N$. The variate for which $\sum_{k\neq j} r_{jk}^2$ or $\sum_{k\neq j} \chi_{jk}^2$ is maximum is therefore the one required. In the conventional (a, b, c, d) notation of a 2 × 2 contingency table this involves calculating, for each pair of attributes, the quantity $(ad - bc)^2/(a + b)(a + c)(b + d)(c + d)$ and summing the resulting symmetrical matrix by rows or columns. If either or both attributes is present or absent in the entire population under study, both numerator and denominator become zero and the resulting indeterminate coefficient is taken as zero.

* *C.S.I.R.O. Computing Research Section, Canberra, A.C.T., Australia.* (Permanent address of Professor Williams: *Botany Department, The University, Southampton.*)

Macnaughton-Smith (1965), in a general review of divisive systems using an information-theory model, similarly concludes that a close numerical approximation to the information-content of an attribute x_j is given by $\sum_{k \neq j} X^2_{jk}$, which for qualitative data is equivalent to the Stein solution. In our original communication we used a factor-analysis model, suggesting that the attribute with maximum $\Sigma|r_{jk}|$ would commonly possess the highest loading on the first averoid axis of the reflected correlation matrix. The KDF9 program uses ΣX^2 as the normal division parameter; the Pegasus, IBM, and Control Data programs use $\Sigma|r|$, i.e. $\Sigma\sqrt{(X^2/N)}$; the standard Elliott program also uses $\Sigma|r|$, but alternative versions are available using ΣX^2 and the two empirical coefficients $\Sigma|ad - bc|$ and $\Sigma(ad - bc)^2$.

Although ΣX^2 undeniably provides the maximum-information split, it has the disadvantage (for some purposes) of fragmenting the analysis by initially splitting off outliers from the population. The less-efficient $\Sigma|r|$ tends in practice to provide a more even split, and therefore to preserve inherent similarities in the early stages of the analysis. Macnaughton-Smith (1965) has pointed out that this is equivalent to foregoing some degree of statistical efficiency in return for partial fulfilment of a utilitarian requirement, and suggests that a formalized decision-function along these lines might provide a better solution. This suggestion is under investigation; meanwhile, in our experience, $\Sigma|r|$ still provides the best general-purpose solution.

Missing values

Only the KDF9 program has this facility; the 2 × 2 table is constructed from the both-attributes-known subset of individuals (say, n out of N) and the resulting X^2 taken as a contribution to ΣX^2 without scaling for reduced N. Provision could be made in the $\Sigma|r|$ programs, providing the proportion of missing values was not very great, by taking $\sqrt{(X^2/n)}$ as the best predictor of $\sqrt{(X^2/N)}$. A theoretically better solution would be to use the partition correlation coefficient of Williams and Dale (1962), taking unknowns as "qualitative," and known binary values as "quantitative"; but the computation is cumbersome and would probably prolong the analysis unduly.

Stopping-rules and hierarchical levels

The populations under study are almost invariably samples; and it follows that, in the later stages of the analysis, spurious divisions arising from sampling error are likely to be encountered. Stenhouse (personal communication) has in fact shown that these programs will subdivide data generated from a set of random numbers. It is, too, common experience that the later stages of an analysis frequently proliferate into a large number of unprofitable divisions. A measure is required which will terminate the analysis before this point, and this measure is further required to serve as a level of the "importance" of each subdivision, and to fall monotonically as the analysis proceeds.

No completely satisfactory stopping-rule is known. Goodall (1953) originally suggested that the proportion of X^2 values in the matrix which exceed significance should be taken as a measure of significance; i.e., the null hypothesis is complete independence, and it is postulated that 1 in 20 of the values would be expected to exceed the $P = 0\cdot05$ level ($X^2 = 3\cdot84$ for one degree of freedom) by chance alone. This method has not been tested on any of the modern programs. The Pegasus, Elliott 803 and Control Data 3600 programs use an empirical and frankly pragmatic measure, the single greatest X^2 encountered in the matrix, printed out with a Yates correction. For Pegasus the terminal value is $3\cdot84$ or $N/32$, where N is the size of the total population; for the Elliott 803 it is $6\cdot63$ (corresponding to $P = 0\cdot01$); for the Control Data 3600 any value can be specified. As a result of the Yates correction, the smallest population capable of generating a given X^2 is, approximately, the nearest integer to $(X^2 + 4)$; and the IBM 704 program (Grunow, 1964) appears to use population size ($N = 50$) as a stopping rule. The KDF9 program uses ΣX^2 itself, assigning to it the probability of a X^2 with as many degrees of freedom as there are attributes in the population. There are statistical objections to this practice, since normal probabilities do not apply to a maximized sample, but it is the least empirical method currently in use.

More rigorous statistical investigation of stopping rules is greatly needed. The difficulties are exacerbated by the fact that the normal strategy—which we believe is used by all programs—is to work down the positive side of the hierarchy, then scan back to the last undivided population, and so on until the extreme negative side is reached. An attractive modification would be to scan evenly down all branches, each time making that division associated with the highest available value of the level-measure in use, until a specified number of groups was obtained. We are planning a version of the Control Data 3600 program along these lines.

2. Computer considerations

Storing the data

The primary data constitute a Boolean array of n individuals specified by the possession or lack of p attributes, and, in view of the size of most practical problems, it is essential that the elements of the array should be held as binary digits. It is therefore necessary to decide whether the data—customarily punched by the user as individuals—should be held in the computer by individuals or by attributes. If it is desired to divide the *individuals* into groups ("normal" analysis) the computer is required to set up all possible 2 × 2 tables between pairs of attributes over all individuals; if it is desired to group *attributes* ("inverse" analysis—Williams and Lambert, 1961) the tables are set up between pairs of individuals over all attributes. No provision for automatic transposition is made in the Pegasus or KDF9 programs, nor, we believe, in that for the IBM 704; in these cases, if both normal and inverse analyses are required, separate data-tapes must be prepared by the user. The Elliott 803 possesses a subsidiary program which punches out a transposed data-tape, and the Control Data 3600 program provides for automatic transposition on entry if required. In all that follows we shall assume that a "normal" analysis is required.

If the data is held by individuals, all n individuals must be passed successively through a series of masks representing all possible pairs of attributes, a total of $\frac{1}{2}np(p-1)$ passes. If held as attributes, however, these can be collated directly, requiring only $\frac{1}{2}p(p-1)$ passes. For a fast program it follows that the data must be held by attributes, but only the Control Data program operates in this manner; it assumes that the data will be supplied by individuals, and transposes on entry if a "normal" analysis is required.

Identification of sub-populations

In the Pegasus program the individuals are sorted—i.e., re-positioned—in the computer store; the identity of the individuals is thereby lost, and the sub-populations are known only by their attribute-structure. The user has in this case commonly prepared a set of edge-punched cards, one for each individual, which are sorted outside the computer into the groups defined by the subdividing attributes. To obviate this, both the Elliott and Control Data programs retain the identity of the individuals and print out, not only the attribute-structure, but also the composition of every "final" group by individuals.

Individuals are usually numbered serially in the order they are read into the computer. This is acceptable if the data are held on tape; if, however, they are on cards, the consequences of a single card out of order would be serious for the user. It is in principle desirable that each individual should be supplied with a tag number; these numbers would be stored as a tag vector, and used to interpret serial numbers before output. Such a system would in addition enable the user to employ tags not in a complete arithmetic sequence. We are contemplating incorporating this facility into later versions of the Control Data program.

It is sometimes necessary to specify, from the set of data presented to the computer, a particular sub-population for analysis; this usually arises when it is wished to continue an analysis which has been interrupted. In the Pegasus and Elliott programs provision is made for analyzing a sub-population of given attribute structure; in the Control Data program an "individual-mask" is required, specifying the serial numbers of the individuals in the desired sub-population.

Attribute masking

The user may ask that only certain specified attributes be used in the analysis. This provision exists on the Pegasus and Control Data programs, but is absent from the KDF9 and the current version of the Elliott program; we have no information concerning the IBM program. More important from the point of view of speed is the existence of an *internal* attribute-masking system. In the later stages of an analysis many attributes, not themselves used in subdivision, may have become sorted between sub-populations, so that in any sub-population under study a number of attributes may be possessed or lacked by all individuals. All X^2 values for these attributes are indeterminate and taken as zero. Now, the setting up of the $\frac{1}{2}p(p-1)$ contingency tables begins with a first pass of $(p-1)$ tables, during which all indeterminate attributes will be encountered; if these are recorded in an internal mask applicable only to the sub-population under study, no tables involving these attributes need be set up again. It is easily shown that, if among a complete set of p attributes there are q which are possessed or lacked by all individuals, the reduction in number of tables to be set up is $\frac{1}{2}q(2p-q-3)$. This can be considerable; if, in a set of 30 attributes, 10 have become indeterminate, the number of tables to be set up is reduced from 435 to 200. This provision is thus necessary if the program is to be fast. It exists in the Pegasus and Control Data programs, is lacking for the KDF9, and we have no information in this regard concerning the IBM program. In the Elliott program the analysis begins with a count of the number of individuals in each attribute; the counts are stored and used as marginal totals in the X^2 calculations. If, however, the count is zero or equal to the size of the sub-population under study, the attributes concerned are recorded as an internal mask, and only $\frac{1}{2}(p-q)(p-q-1)$ tables are set up.

3. Specification of the Control Data 3600 program

Input

The first card read in contains four quantities, n, p, T and L, where n is the number of individuals, p is the number of attributes, T is unity if the data is to be read in transposed (for a "normal" analysis) or zero if it is to be transposed, and L is the level of subdivision ($3\cdot84$ for $P = 0\cdot05$).

Then follows the data. Two types of input are allowed.

(i) If the first individual contains species 1, 3, 5 to 8 (inclusive), and 10 the first card would be punched as 1/3/5–8/10. This is a useful input method when the total number of attributes possible may be large but each individual contains only a few.

(ii) Alternatively, the above individual could be punched as 5364.; that is, in an octal form.

The form of data must be the same for a given run of the program and in both cases the *last card* has an * immediately following the stop. Blanks may be inserted anywhere, for convenience of layout, and continuation cards are allowed but data must not be punched in columns 73–80. These columns may be used for identification purposes.

A mask card follows the data; it specifies, in the case of a normal analysis, which attributes are to be used so a subset can be chosen. It is punched in the same way as the data cards. Thus if attributes 1–10 and 20–29 only are to be considered, the mask card would be punched 1–10/20–29.*

Method

The program first reads the control card, the data and the attribute mask. Using the value of n so obtained it forms the "individual" mask which describes the contents of the first group. The 2×2 tables are then set up between those attributes contained in the "attribute" mask. From these tables the coefficients are computed and the summations indicated in Section 1 are formed. The "most-significant" attribute is thus determined and the first subdivision determined.

At this stage there are two "individual" masks; one is stored away for later consideration and the other used as before. After all data have been passed through this mask, the number of individuals in each data item is compared with the number in the mask itself. If these are equal then the attribute concerned is removed from the "attribute" mask because it will contribute only zero coefficients. This subdivision process is continued until the stopping-rule is satisfied, and then the program returns to deal with "individual" masks previously stored away. When none of these remain, the calculation is at an end. During the course of the subdivisions, certain information about the construction of the hierarchical table is stored so that it may be plotted at the end.

Output

The first information printed out is a copy of the data. Method (i) above is always used for output, even though the data was fed in in octal. However, if the analysis is normal, the data will be in the transposed form, that is, the first line will show which individuals contain attribute number one.

At the top of the next page the "attribute" mask is printed to indicate which attributes are used in the analysis. (This is always a copy of the last card read.) Then follows:

(i) "Number of individuals in this group is X."
(ii) A list of the X individuals.
(iii) "Number of chisquareds above L is Y." (This L is the level L specified on the control card.)
(iv) "Maximum chisquared value is V, the sort is on attribute number N."

After this, the items (i)–(iv) are repeated unless the group is "final," in which case "Final" replaces (iii) and (iv).

The last thing the program does is to plot the hierarchical table on the graph plotter. The form of the table is the same as those shown as Figs. 4 or 5 in Williams and Lambert (1961).

Acknowledgements

Our thanks are due to Dr. B. R. Ivemy-Cook for information concerning the Elliott program, and to Mr. A. I. Jones for the details of his KDF9 program.

References

DALE, M. B. (1964). "The application of multivariate methods to heterogeneous data," *Ph.D Thesis, Southampton.*

GOODALL, D. W. (1953). "Objective methods for the classification of vegetation. I," *Aust. J. Bot.*, Vol. 1, p. 39.

GRUNOW, J. O. (1964). "Objective classification of plant communities," *S. Afr. J. Agric. Sci.*, Vol. 7, p. 171.

MACNAUGHTON-SMITH, P. (1965). *Some statistical and other numerical techniques for classifying individuals*, H.M.S.O. (*in the press*).

WILLIAMS, W. T., and DALE, M. B. (1962). "Partition correlation matrices for heterogeneous quantitative data," *Nature*, Vol. 196, p. 602.

WILLIAMS, W. T., and LAMBERT, J. M. (1959). "Multivariate methods in plant ecology, I," *J. Ecol.*, Vol. 47, p. 83.

WILLIAMS, W. T., and LAMBERT, J. M. (1960). "Multivariate methods in plant ecology, II," *J. Ecol.*, Vol. 48, p. 689.

WILLIAMS, W. T., and LAMBERT, J. M. (1961). "Multivariate methods in plant ecology, III," *J. Ecol.*, Vol. 49, p. 717.

WILLIAMS, W. T., and LANCE, G. N. (1958). "Automatic subdivision of associated populations," *Nature*, Vol. 182, p. 1755.