Note on a new information–statistic classificatory program

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

This note describes a new divisive monothetic method for the classification of data specified by binary attributes. Its advantages over an earlier method are outlined.

(First received November 1967)

In a previous communication (Lance and Williams, 1965) we have examined the properties of the classificatory system known as 'association analysis' (our program ASSO). Given a population defined by binary attributes, this program seeks that attribute which provides the best overall prediction of the joint variance of the remainder, and divides the population into those possessing, and those lacking, the chosen attribute; the process is repeated on the resulting sub-populations, or 'groups'. It has been pointed out by Macnaughton-Smith (1965) that the process also defines a near-optimum information split; and in view of the successes obtained with an information statistic in the corresponding agglomerative system (Lambert and Williams, 1966), there are obvious attractions in the possibility of an information-statistic counterpart of ASSO.

Given a population of n elements defined by s binary attributes, such that the jth attribute is present in a_j elements, we define an information content of the population, I, such that

$$I = sn \log n - \sum_{j=1}^{s} \{a_j \log a_j + (n - a_j) \log (n - a_j)\}.$$

Alternative derivations of this statistic are given in Lance and Williams (1966) and in Macnaughton-Smith (1965). If a population (i) is divided into two groups (g) and (h), we define an information-fall, $\Delta I_{(gh,\,i)}$ by the relation

$$\Delta I_{(gh,\,i)} = I_i - I_g - I_h.$$

The population is dichotomised on each attribute in turn, the ΔI for each division calculated, and division effected on that attribute for which ΔI is maximum. The resulting hierarchy is plotted with the total I values as hierarchical levels and, since these by definition fall monotonically, 'reversals' in hierarchical level (which are troublesome, and not uncommon in ASSO) are impossible.

Stopping-rules are available, which in this case serve primarily to define the group next to be divided. In the terminology of Lambert and Williams (1966), a Type 1 stopping-rule selects that group whose $\Delta I_{\max}$ is the highest of all the remaining group maxima; this involves notionally dividing every remaining group. The corresponding Type 2 rule selects that group whose I is greatest, and obviously requires fewer operations. In our program termination is at a specified *number* of groups. Alternatively, with a Type 1 rule, advantage could be taken of the fact that $2\Delta I$ is a biased estimate of χ^2 with as many degrees of freedom as there are attributes, less one d.f. for each attribute used for division up to that point. Theoretically, in a sub-population of n' members with s' determinate attributes remaining, $2I$ itself is an estimate of a χ^2 with $s'(n' - 1)$ degrees of freedom; but in most cases this number is very large, and the test of significance correspondingly weak.

The absence of reversals and the simple and rigorous stopping-rules would alone make the system attractive by comparison with association analysis; but if a Type 2 stopping-rule (which we normally recommend) is in use it has the added advantage of computational speed. As a comparative example, we have taken the population of 18 elements specified by 818 binary attributes examined in Webb *et al.* (1967); this was divided into 6 groups by association-analysis, the computation requiring 23 minutes 28 seconds. The computation on the new system with a Type 2 stopping-rule produced the identical answer in 10 minutes 47 seconds.

The program (entitled DIVINF) has been written for the Control Data 3600 computer at Canberra, and listings are available on request.

References

LAMBERT, J. M., and WILLIAMS, W. T. (1966). Multivariate methods in plant ecology. VI. *J. Ecol.*, Vol. 54, p. 635.

LANCE, G. N., and WILLIAMS, W. T. (1965). Computer programs for monothetic classification ('Association analysis'), *Computer Journal*, Vol. 8, p. 246.

LANCE, G. N., and WILLIAMS, W. T. (1966). Computer programs for hierarchical polythetic classification ('Similarity analyses'), *Computer Journal*, Vol. 9, p. 60.

MACNAUGHTON-SMITH, P. (1965). *Some statistical and other numerical techniques for classifying individuals*, H.M.S.O., Home Office Research Unit Report, No. 6.

WEBB, L. J., TRACEY, J. G., WILLIAMS, W. T., and LANCE, G. N. (1967). Studies in the numerical analysis of complex rain-forest communities. I. *J. Ecol.*, Vol. 55, p. 171.

* *C.S.I.R.O. Division of Computing Research, Canberra, A.C.T., Australia.*