62H30

BULL. AUSTRAL. MATH. SOC.
VOL. 32 (1985), 473-475.

CLUSTER ANALYSIS VIA NORMAL MIXTURE MODELS

KAYE ENID BASFORD

The technique of clustering uses the measurements on a set of elements to identify clusters or groups of elements, such that there is relative homogeneity within the groups and heterogeneity between the groups. Under the mixture maximum likelihood approach to clustering, the elements are assumed to be a sample from a mixture of various proportions of a specified number of populations, By adopting some parametric form for the density function in each underlying population, a likelihood can be formed in terms of the mixture density and the unknown parameters estimated using the likelihood principle. The allocation of the elements to the populations is determined on the basis of the estimated posterior probabilities of population membership.

Various aspects of this approach to clustering are examined in the case where the component populations are assumed to have multivariate normal distributions. Initially, the problems associated with likelihood estimation in this context are explored. Unfortunately with mixture models, the likelihood equation usually has multiple roots, so there is the question of which root to choose. In the case of equal covariance matrices the situation is straightforward, in that the maximum likelihood estimator exists and is consistent. An example is presented, however, to demonstrate that the adoption of a homoscedastic normal model, in the presence

Received 29 July 1985. Thesis submitted to University of Queensland, February 1985. Degree approved July 1985. Supervisors: Assoc. Prof. G.J. McLachlan, Dr. D.E. Byth.

https://doi.org/10.1017/S0004972700002586 Published online by Cambridge University Press

of some heteroscedasticity, can considerably influence the likelihood estimates, in particular, of the mixing proportions, and therefore the consequent clustering of the sample at hand.

With the mixture approach to clustering, each element in the sample is allocated to one of the component populations on the basis of the estimated posterior probabilities of population membership. Consideration is given to assessing the performance of the mixture approach by averaging appropriate functions of these posterior probabilities. In the case where the superpopulation does consist of individual component populations in number equal to that specified with the particular application of this method, these functions can be regarded as estimates of the correct allocation rates for the individual populations, as well as for the overall mixture. As the proposed method of estimation can produce biased estimates, the bootstrap procedure is studied for its effectiveness in reducing this bias. Three real data sets are considered as well as a detailed simulation study . It is shown that the proposed estimates generally provide useful information on the unobservable allocation rates of the mixture approach. Encouraging results are obtained for the bootstrap method of bias correction applied to the estimates of the individual and overall allocation rates.

The role of the mixture approach is investigated, also, for the clustering of treatments from a randomized complete block design. In this context, it provides a concise way of summarizing differences amongst the treatments. It is shown that the implementation of this technique is straightforward for fixed, but not random block effects. The difficulty is that with the latter model the amount of computation becomes prohibitive for even a moderate number of treatments. This approach is illustrated by the application of the mixture method to three data sets from randomized complete block designs. The first two examples concern real data already used in the literature to illustrate other techniques for grouping means under a fixed effects model. The third example consists of data simulated according to

https://doi.org/10.1017/S0004972700002586 Published online by Cambridge University Press

a randomized complete block design with random block effects, where the true grouping is known and where the number of treatments is small enough for the mixture method to be implemented under this mixed model. The results under this model are compared with those obtained by treating the block effects as fixed, in an instance where they are actually random.

Most available clustering techniques are applicable only to a two-way data set, where one of the modes (the elements) is being partitioned into groups on the basis of the other mode (the measurements). If, however, the data set is defined as a three-way array, then a multivariate technique, which will cluster the elements (one of the modes) on the basis of the total information available (the other two modes simultaneously), is required. It is shown that by appropriate specification of the underlying model, the mixture method of clustering can be applied`in this context. This is illustrated using soybean data, which consist of multi-attribute measurements on a number of genotypes, each grown in several environments. This approach is particularly useful here, as the genotype by environment interaction is able to be incorporated directly into the underlying model. Although the problem is set in the framework of clustering genotypes, the technique is applicable to other types of three-way data, and so is of considerable potential.

Department of Agriculture,
University of Queensland,
St. Lucia, Q. 4067,
Australia.

https://doi.org/10.1017/S0004972700002586 Published online by Cambridge University Press