BULL. AUSTRAL. MATH. SOC.
VOL. 24 (1981), 309-312.

62H30

# On the mixture maximum likelihood approach to estimation and clustering

Selvanayagam Ganesalingam

In this study attention is focussed on the performance of the mixture maximum likelihood approach as an estimation and a clustering procedure. There is available a sample of $p$-dimensional observations where each may belong to one of several subpopulations. Estimation on the basis of this sample is considered in a cluster analysis context where it is not known from which subpopulation an observation comes. Under the mixture likelihood approach the sample is assumed to have been drawn from a mixture of a specified number of subpopulations in varying proportions. Adopting some parametric form for the density function in each underlying subpopulation, a likelihood can be formed in terms of the mixutre density, and the unknown parameters estimated by maximum likelihood. Parameters of particular interest are the discriminant function coefficients as we shall be primarily concerned with the ability of the mixutre maximum likelihood approach to provide a satisfactory discrimination rule for allocating the initial unclassified observations as well as any subsequent unclassified data.

A comprehensive account of existing work on the estimation problem for finite mixtures of distributions is presented. There are many clustering procedures available, and a review is undertaken of the general classification problem in order to demonstrate where the mixture maximum likelihood approach fits into the existing framework of cluster analysis.

The mixture maximum likelihood approach can be applied, at least in

Received 15 June 1981. Thesis submitted to the University of Queensland December 1980. Degree approved, May 1981. Supervisor: Dr G.J. McLachlan.

principle, to any number of subpopulations with arbitrary distributions, and so the associated computations using the EM algorithm are discussed in a general context. However, the application of the mixture approach is not always straightforward in practice. For example, for a mixture of normal subpopulations with unequal covariance matrices, there are problems with the occurrence of singularities in the likelihood on the edge of the parameter space. Hence the mathematical difficulties associated with a study of the properties of the mixture maximum likelihood approach to clustering are formidable. In an attempt to provide more information on this approach, its properties are investigated in detail for the case of two normal subpopulations with a common covariance matrix. This model would appear to be essential for a substantial mathematical investigation to be undertaken for an arbitrary number of dimensions $p$. Indeed, the problem of providing a meaningful analytical study for more than two subpopulations would appear to be intractable. For a common covariance matrix regularity conditions hold, and so the asymptotic theory of maximum likelihood is used to derive the large sample distributions of the quantities under investigation.

The asymptotic efficiency of the mixture maximum likelihood approach relative to the standard approach, where the classification of the initial sample is known, is derived. Also, the small sample behaviour of the mixture maximum likelihood approach is investigated and it is shown that, although the discriminant function coefficients obtained by the mixture approach may not be reliable, the proportions in which they are estimated are of the order such that the mixture approach performs not too far short of the standard approach at least in terms of the overall expected error rate.

With any approach to clustering there is the problem of assessing its performance. Consideration is given here to the problem of estimating the error rate associated with the discriminant function formed from an unclassified sample by the mixture maximum likelihood approach. The so-called posterior probability based estimator, $e$, formed by averaging the minimum of the posterior probabilities over a set of initial or additional observations (which need not be classified) is introduced for this problem. The bias of $e$ is examined by deriving an asymptotic approximation which indicates that $e$ generally underestimates the true error rate, and this

is subsequently confirmed by a simulative study. The behaviour of $e$ is compared with that of a plug-in type estimator which would appear to be about the only other available estimator for this problem. It is concluded that their relative performance is very similar in terms of bias and mean absolute error. Unfortunately, the magnitude of the bias of these two estimates can be quite large, and so they have to be corrected for bias. This aspect is investigated using the bootstrap and jackknife procedures for bias correction.

The performance of the mixture maximum likelihood approach is compared with another commonly used clustering method, the so-called classification maximum likelihood approach, under both mixture and separate sampling schemes. Under the latter scheme observations have been sampled separately from each subpopulation instead of from a mixture of the underlying subpopulations. It is concluded for mixing propositions which are far from equal under mixture sampling or for disparate sample sizes under separate sampling that the mixture maximum likelihood approach generally has a smaller overall error rate than the classification approach in allocating the initial sample and any subsequent observations. The latter approach appears to have a smaller error rate if the unclassified observations occur in the sample in approximately the same proportion from each subpopulation.

Another important problem investigated in this thesis is the estimation of the mixing proportion $\pi_i$ $(i = 1, 2)$ using the available unclassified data in addition to some classified observations taken separately from each subpopulation. This model corresponds to a number of important examples in practice, especially in the context of the remote sensing of data. In the absence of any classified data the obvious choice of an estimator for $\pi_1$ is the maximum likelihood estimator, $\pi_{1M}$, obtained by the mixture approach. With some classified data available a discriminant function can be formed and applied to the unclasified data to obtain the proportion assigned to the first subpopulation which, after an appropriate correction for bias, yields a useful estimator, $\pi_{1D}$, of $\pi_1$. A derivation is undertaken of the asymptotic efficiency of this easily computed estimator relative to the maximum likelihood estimator, $\pi_{1M}$, based on the unclassified and classified data combined. It is concluded

that if $\pi_{1D}$ suggests that the mixing proportions are disparate then it is worthwhile to proceed further and to compute the asymptotically efficient estimator $\pi_{1M}$, particularly if the sample contains a high proportion of unclassified observations - not an uncommon occurrence in practice since there is generally only a limited number of classified observations available.

A case study is presented to illustrate the use of the mixture maximum likelihood approach as a clustering procedure. Estimation of the error rate associated with this particular application of the mixture approach is undertaken. Also, the performance of the mixture approach is contrasted with that of the classification maximum likelihood method.

119 Kachcheri-Nallur Road,
Nallur,
Jaffna,
Sri Lanka.

https://doi.org/10.1017/S000497270000767X Published online by Cambridge University Press