An alternative proposition to visualize clusters in a MBC context

Benjamin Auder and Gilles Celeux

1 Introduction

This note is devoted to propose a general way to visualize a clustering derived from a mixture model in any context. It is different from the proposal of Biernacki and Marbac (2017). It aims to be clearer, simpler, easier to be computed and more relevant. It simply makes use of the conditional probabilities t_{ik} for an observation i, $i = 1, \ldots, n$ to belong to cluster k, $k = 1, \ldots, K$. It consists essentially of a representation by multidimensional scaling (MDS) of the mixture component distributions.

2 The proposition

From the K conditional distributions of belonging to one of the mixture components knowing the observations, it is possible to compute the symmetrized Kullback-Leibler distance between each couple of components. For $(k, k') \in 1, \ldots, K^2$, it is

$$d_{kk'} = \frac{1}{2}\left[\sum_i \frac{t_{ik}}{t_{ik}+t_{ik'}} \log\frac{t_{ik}}{t_{ik'}} + \sum_i \frac{t_{ik'}}{t_{ik}+t_{ik'}} \log\frac{t_{ik'}}{t_{ik}}\right]. \tag{1}$$

From this table of distances, it is easy to get a representation in $\mathbf{R}^2$ of the clusters $C_1, \ldots, C_K$ by MDS, using an efficient MDS function in R. Obviously, each cluster k receives the weight π_k, namely the mixing proportion of the component k , when running the MDS function.

Some numerical problems can appear in the computation of the distances $d_{kk'}$ when (i) $t_{ik} \approx 0$ and when (ii) both t_{ik} and $t_{ik'} \approx 0$. The case (i) can be answered by imposing a lower bound, say ϵ to t_{ik}. I think that this lower bound could be associated to the smaller non-zero conditional probability that can be numerically computed. (Clearly, this point has to be made more precise.) The case (ii) can be answered by putting $\frac{t_{ik}}{t_{ik}+t_{ik'}} \log\frac{t_{ik}}{t_{ik'}} = 0$.

It is also possible to represent the observations on the MDS graph as illustrative elements on the map representing the clusters as K points $C_1, \ldots, C_K$. (There is in R a MDS function which allows to represent illustrative elements on the computed maps.) In that purpose, there is the need to define the distances between an observation i, $i = 1, \ldots, n$ and the clusters $C_k, k = 1, \ldots, K$. These distances could be defined as follows

$$d_{ik} = \frac{1}{K-1}\sum_{k'} \frac{1}{2}\left[\frac{t_{ik}}{t_{ik}+t_{ik'}} \log\frac{t_{ik}}{t_{ik'}} + \frac{t_{ik'}}{t_{ik}+t_{ik'}} \log\frac{t_{ik'}}{t_{ik}}\right], \tag{2}$$

namely the mean of the contribution of i to the distance $d_{kk'}, k' = 1, \ldots, K$ and $k' \neq k$.

But this way of defining d_{ik} is somewhat cumbersome and the first experiments are not encouraging at all. Thus we propose an alternative and simpler way of defining d_{ik}. First notice that (1) can be written

$$d_{kk'} = \sum_{i=1}^{n} d_{kk'}^{i} \tag{3}$$

with

$$d_{kk'}^{i} = \frac{1}{2}\left[\frac{t_{ik}}{t_{ik}+t_{ik'}} \log \frac{t_{ik}}{t_{ik'}} + \frac{t_{ik'}}{t_{ik}+t_{ik'}} \log \frac{t_{ik'}}{t_{ik}}\right].$$

The key is to define d_{ik} as $d_{kk'}^{i}$, with $k' = i$ and thus with $t_{ik'} = 1$. It leads to

$$d_{ik} = \frac{1}{2} \frac{t_{ik}-1}{1+t_{ik}} \log t_{ik}. \tag{4}$$

3 The suggested representation

We think that a good way to represent the clusters and the observations on the MDS map could be to represent the clusters $C_1, \ldots, C_K$ with big points of different colours and the observations $i, i = 1, \ldots, n$ with smaller points of the same color than the cluster $k(i)$ such that $k(i) = \max_k t_{ik}$.

No doubtful and maybe misleading confidence regions or levels, but the mixing of colours will indicate the degree of component overlap. An last but not least, such graphical representations will take advantage from the contribution indices of factor analysis.

Reference

[1] Biernacki, C., Marbac, M. (2017) Gaussian based visualization of Gaussian and non-Gaussian model-based clustering. Working paper.