Forming Clusters from Census Areas with Similar Tabular Statistics

Murray Jorgensen

Department of Statistics, University of Waikato, Hamilton, New Zealand

National statistical offices are beginning to make available tabular data on census variables for for fairly small geographical regions. For example the United Kingdom National Statistics Office has a web site `http://www.neighbourhood.statistics.gov.uk/` from which tables at the post code level may be obtained. Similarly Statistics New Zealand provides regional tables through its web site
`http://www.stats.govt.nz/products-and-services/table-builder/default.htm`.
If it is possible to group the small areas for which data is available into larger regions such that the cell proportions in the tables for areas within a region are similar this will be of great value for an exploratory analysis of the data.

Consider clustering observations where each observation is a table consisting of counts in each of a number of categories. As an example consider Age group by Sex data from the New Zealand Census of 2006. This yields about fifteen hundred 2×15 tables, one for each 'area unit'. The observed category counts for a particular area unit may be regarded as a sample from a 30-category multinomial distribution.

We adopt the following model for the probability $p_i(\boldsymbol{y}_i)$ that the ith area unit has resident counts of $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ip})'$ in the $p = 30$ demographic categories

$$\begin{aligned} p_i(\boldsymbol{y}_i) &= p_i(y_{i1}, y_{i2}, \ldots, y_{ip}) \\ &= \sum_{j=1}^{q} \pi_j \frac{m_i!}{y_{i1}! y_{i2}! \cdots y_{ip}!} \lambda_{j1}^{y_{i1}} \lambda_{j2}^{y_{i2}} \cdots \lambda_{jp}^{y_{ip}} \\ &= \sum_{j=1}^{q} \pi_j p_{ij}(\boldsymbol{y}_i) \end{aligned}$$

where m_i stands for $y_{i1} + y_{i2} + \cdots + y_{ip}$

This is essentially a finite mixture of multinomial distributions Multi$(m_i, \lambda_{j1}, \lambda_{j2}, \ldots, \lambda_{jp})$, except that the numbers m_i may vary depending on the area unit i. This model allows us to utilise model-based clustering [2] to group the area units into q clusters.

Grouping the areas into a moderate number of clusters with differing patterns of cell proportions provides a way to understand the "message" of the data, especially where the names and locations of the areas are known to the analyst. The same methodology was used by Jorgensen [1] to cluster packet size distributions in packet flows over the internet between computers. That analysis was not as easy to interpret in the absence of background information about the various different packet flows.

References

[1] M. A. Jorgensen. Using multinomial mixture models to cluster internet traffic. *Austral. and New Zealand J. Statist.*, 46:205–218, 2004.

[2] G. J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, New York, 2000.