Clustering Multivariate Categorical Data Using Exact ICL Method

Manasi Mohan Narsapur, Master of Science in Computer Science

University of Dublin, Trinity College, 2020

Supervisor: Dr. Arthur White

Clustering, an integral part of data analysis, is a process of grouping together observations based on characteristics, which aids discovery of hidden information from a dataset. The latent class analysis (LCA) modelling algorithm is administered with the exact integrated complete likelihood (ICL) method, to build an algorithm named *exactICLforLCA* in this dissertation with the intent to cluster the multivariate categorical data utilizing the exact ICL method. Concepts such as Bayesian inference, beta distribution, Dirichlet distribution were applied to produce the notation obtained by administering ICL method on LCA modelling algorithm which was used to gauge the improvement in clustering. The algorithm was fit by utilizing real world data such as Alzheimer's dataset and simulated datasets. The results illustrated that the algorithm successfully assists in the improvement of clustering for a dataset by identification of the optimum number of groups, affirmed by the increase in the ICL value. The new matrices generated by the *exactICLforLCA* algorithm displayed an improvement in clustering and the average increase in the ICL value for simulated datasets was approx. 140.77 and 65.307 values for real data when compared to the fit by expectation–maximization algorithm for an experiment run for 20 iterations for both the datasets. Data clustered with the help of expectation–maximization algorithm were compared against the fit by the *exactICLforLCA* algorithm with the help of a cluster evaluation technique called `randIndex` to calculate the similarity among them.

Keywords— integrated complete likelihood, latent class analysis, Bayesian inference, cluster analysis