Abstract No: PO-33

A comparison of distance-based and model-based clustering methods

H. A. D. D. Nadeekantha[1*], H. W. B. Kavinga[1], A. Gunawardana[2] and D. M. P. V. Dissanayaka[1]

[1] Department of Statistics & Computer Science, University of Kelaniya, Sri Lanka
[2] Department of Decision Sciences, University of Moratuwa, Sri Lanka
damika.nadeekantha@gmail.com*

Most of the statistical techniques assume the homogeneity of the sample data. However, not all the time, real-world samples are homogeneous. The existence of subgroups within a population leads to the non-homogeneity of the sample. In this case, it is not accurate to model the population using a single probability distribution. Hence it is essential to check the homogeneity of the sample. Clustering, an unsupervised learning technique, is being used to discover a population's subgroups and group each observation into a specific cluster. Mainly, clustering algorithms can be divided into two groups, namely model-based and distance-based algorithms. Model-based algorithms assume a probability distribution for clustering, while distance-based algorithms use a distance metric to classify observations into clusters. In the literature, it was suggested that the model-based clustering methods perform better than the distance-based methods using summary statistics and visualizations. In this study, an inference-based procedure has been used to assess the above claim. To compare the performances of model-based and distance-based algorithms, an extensive simulation study was conducted. In the simulation study, two univariate Gaussian mixtures with different parameter settings (mean, standard deviation, and sample size) were combined to generate a non-homogeneous sample. Then, model-based and distance-based algorithms were applied to the same simulated datasets with different cluster structures, knowing the actual cluster memberships. Further, the effect of bimodality conditions of Gaussian mixtures on both clustering methods was checked. To assess the performance of the two methods, identifying the correct number of clusters, Cluster Identification Ability (CIA), and categorizing the observations into the correct cluster memberships (clustering accuracy) were computed. CIA was computed using the percentage of iterations that identified the correct number of clusters, and clustering accuracy was measured using the Adjusted Rand Index (ARI). For most of the simulation settings, both methods required a sample size of less than 200 to achieve high clustering accuracy (approximately mean ARI value of 0.8). For example, a simulation setting with a mean difference of 3.1 and a standard deviation of 0.5 required sample sizes 20 and 10 for the model-based and distance-based methods, respectively. These minimum sample sizes vary depending on the method's high clustering accuracy, and in some cases, those are approximately the same. The inference-based study which is performed using the paired Wilcoxon signed-rank test indicated that the claim "model-based method outperforms distance-based method, or both performs similarly" is valid 82.7% of the time at a 5% level of significance. In conclusion, the CIA and clustering ability of the model-based method increased with the increment of sample size when the bimodality conditions were satisfied by the mixture. For the distance-based method, both abilities decreased as the sample size increased when the bimodality conditions were not satisfied by the sample.

Keywords: Non-homogeneity, Clustering, Model-based, Distance-based, Gaussian-mixtures.