Supplemental Information

BASIC DESCRIPTION OF CLUSTER ANALYSIS METHODS

K-means clustering plots patients in *N*-dimensional space, in which *N* is the number of data elements describing each patient. The algorithm then partitions patients into a number of clusters chosen by the analyst, such that each patient is closer to the central tendency of their cluster than they are to the central tendency of the other clusters, while making the central tendency of the developed clusters as far apart as possible. Thought of another way, clusters are developed by testing each possible combination until the partition is found in which the data points within each cluster are maximally similar and the cluster averages are maximally different between clusters.[40]

The Gower distance, or Gower similarity, is a method for assessing the distance between data points. This can also be thought of as the degree to which 2 data points are similar or different. The measuring of distance between data points is complicated when using a data set with mixed continuous or interval data (age) and dichotomous data (presence or absence of RH). For categorical values, the distance will either be 1 or 0. For continuous or interval data, the distance can be much larger. In our sample, the difference in age in months could be as high as 36. Gower method divides the actual distance between 2 continuous variables by the range of those variables, preventing continuous variables from becoming too influential in a cluster partition.[41]

As noted above, in K-means partition, the number of clusters is specified by the analyst before the computer developing the partition. We specified 2 through 10 clusters. This leaves the task of choosing which partition, with what cluster number, is optimal. The resulting partitions were evaluated by 2 methods to determine the optimal number of clusters: silhouette width and gap statistic. Silhouette width describes the degree to which the mean distance between cluster members and the central cluster tendency is less than the distance between the central cluster tendencies. Silhouette width varies from −1 to 1, with the optimal cluster result being the one with the maximal silhouette width.[39,42] Gap statistic looks at the gap between the dispersion of members within a cluster and the dispersion resulting from a random uniform distribution of the members. The optimal cluster result is felt to be the smallest number of clusters (k) such that the gap statistic for k+1 clusters is maximized but within 1 SD of the gap statistic for k clusters.[43]

Agglomerative hierarchical clustering starts with each patient as a cluster of 1 and then serially combines clusters that are the nearest to one another based on the Gower distance between the clusters. There are a number of ways of computing this distance, and we have chosen to use the complete or “farthest neighbor” method, in which the intercluster distance is the distance between the 2 individual cluster members that are the furthest apart from one another. Clusters are combined, iteratively, until the entire database is agglomerated into 1 cluster.[44]

Divisive hierarchical clustering is the inverse of agglomerative clustering. The entire data set is divided into 2 clusters so as to maximize the Gower distance between them. Again, the complete method was used for determining distance. Resulting clusters are then divided iteratively until each patient stands alone as a cluster of 1. For hierarchical clustering, the entire hierarchical tree is the product; there is no optimal cluster number. The analyst is free to look at whatever partition best suits their research question.[45]

SUPPLEMENTAL REFERENCES

40. 365 Data Science. What is K-Means clustering? Available at: https://365datascience.com/tutorials/python-tutorials/k-means-clustering/. Accessed August 20, 2021

41. McCaffrey J. Example of calculating the Gower distance. Available at: https://jamesmccaffrey.wordpress.com/2020/04/21/example-of-calculating-the-gower-distance/.

42. 42. Dalwani K. Using silhouette analysis for selecting the number of cluster for K-Means clustering (part 2). Available at: https://kapilddatascience.wordpress.com/2015/11/10/using-silhouette-analysis-for-selecting-the-number-of-cluster-for-k-means-clustering/. Accessed August 20, 2021

43. Lohr T. K-Means clustering and the gap-statistic. Available at: https://towardsdatascience.com/k-means-clustering-and-the-gap-statistics-4c5d414acd29. Accessed August 20, 2021

44. Data Novia. Agglomerative hierarchical clustering. Available at: https://www.datanovia.com/en/lessons/agglomerative-hierarchical-clustering/. Accessed August 20, 2021

45. Data Novia. Divisive hierarchical clustering. Available at: https://www.datanovia.com/en/lessons/divisive-hierarchical-clustering/. Accessed August 20, 2021