Israel Nolazco

israelnolazco@lewis.edu

A Summary of Cluster Analysis

August 25, 2020

Table of Contents

1. **What is it?**
2. **A downside?**
3. **Caution**
4. **Measuring Similarity**
5. **Hierarchical Cluster Procedures**
6. **Deriving Non-Hierarchical Clusters**
7. **DERIVING THE FINAL CLUSTER SOLUTION**
8. **INTERPRETING, PROFILING AND  VALIDATING  CLUSTERS**
9. **What is it?**

Cluster analysis is a technique in which data points are group together based on their patterns on variation.  The way these patterns are group is by distance from one another.  This ensures that the groups have data points which similar characters from one another.  It has been referred to as Q analysis, typology construction, classification analysis, and numerical taxonomy.  In short, the essence of all clustering approaches is the classification of data as suggested by “natural” groupings of the data themselves.

1. **A downside?**

Believe it or not, cluster analysis can be criticized for working too well in the sense that statistical results are produced even when a logical basis for the cluster is not apparent.  Therefore, the analyst should have a strong conceptual basis to deal with issues such as why groups exist in the first place and what variables logically explain why objects end up in the group that they do. In short, you have to be some sort of an expert in the subject matter.

1. **CAUTION**

* CLUSTER ANALYSIS IS DESCRIPTIVE, ATHEORITEICAL, AND NONINFERENTIAL.   There is no statistical basis upon which to draw inferences from a sample to a population and it is primarily used for an exploratory technique
* CLUSTER ANALYSIS WILL ALWAYS CREATE CLUSTERS, REGARDLESS OF ACTUAL EXISTENCE OF ANY STRUCTURE IN THE DATA.  Remember that just because a cluster can be found does not validate their existence.

1. **Measuring Similarity**

Interobject similarity is an empirical measure of correspondence, or resemblance, between objects to be clustered.  It  can be measured in a variety of ways, but three methods dominate the applications of cluster analysis:

• Correlational Measures - The inter-object measure of similarity that probably comes to mind first is the correlation coefficient between a pair of objects measure on several variables.

• Distance Measures - Euclidean Distance, Squared Euclidean Distance, City-block Distance, Chebychev Distance, and Mahalanobis distance (D2).

• Association - measures of similarity are used to compare objects whose characteristics are measured only in nonmetric terms

1. **Hierarchical Cluster Procedures**

Hierarchical procedures involve a series of n-1 clustering decisions (where n equals the number of observations) that combine observations into a hierarchy or a treelike structure.

Agglomerative Methods - each object or observation starts out as its own cluster an is successively joined, two most similar cluster at a time until only a single cluster remains.

Divisive Methods - all observations tart in a single cluster and are successively divided (first into two clusters, then three, and so forth) until each is a single-member cluster.

1. **Deriving Non-Hierarchical Clusters**

•Nonhierarchical clustering methods require that the number of clusters is specified before assigning observations:

–The sequential threshold method assigns observations to the closest cluster, but observation cannot be re-assigned to another cluster following its original assignment.

–Optimizing procedures allow for re-assignment of observations based on the sequential proximity of observations to clusters formed during the clustering process.

1. **DERIVING THE FINAL CLUSTER SOLUTION**

•There is no single objective procedure to determine the ‘correct’ number of clusters.  Rather the researcher must evaluate alternative cluster solutions on the following considerations to select the “best” solution:

- Single-member or extremely small clusters are generally not acceptable and should generally be eliminated.

- For hierarchical methods, ad hoc stopping rules, based on the rate of change in a total similarity measure as the number of clusters increases or decreases, are an indication of the number of clusters.

- All clusters should be significantly different across the set of clustering variables.

- Cluster solutions ultimately must have theoretical validity assess through external validation.

1. **INTERPRETING, PROFILING AND  VALIDATING  CLUSTERS**

•The cluster centroid, a mean profile of the cluster on each clustering variable, is particularly useful in the interpretation stage.

- Interpretation involves examining the distinguishing characteristics of each cluster’s profile and identifying  substantial differences between clusters

- Cluster solutions failing to show substantial variation indicate other cluster solutions should be examined.

The cluster centroid should also be assessed for correspondence with the researcher’s prior expectations based on theory or practical experience.

•Validation is essential in cluster analysis since the clusters are descriptive of structure and require additional support for their relevance:

- Cross-validation empirically validates a cluster solution by creating two sub-samples (randomly splitting the sample) and then comparing the two cluster solutions for consistency with respect to the number of clusters and the cluster profiles.

- Validation is also achieved by examining differences on variables not included in the cluster analysis but for which there is a theoretical and relevant reason to expect variation across the clusters.